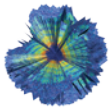
THIS WEEK

EDITORIALS

PUBLISHING Oldest journal highlights age-old struggle for reason **p.130**

WORLD VIEW Battle for Earth science in US schools needs you **p.131**



IMAGING X-rays reveal 3D structure of tiny virus particle **p.133**

An array of problems

Political interference in the selection process for the headquarters of the Square Kilometre Array should not go unchallenged.

When David Cameron addressed Australia's Parliament last November, the British prime minister referred briefly to the Square Kilometre Array (SKA), "the world's largest radio telescope". The project's headquarters, he noted, were in Manchester, UK. Not so. The location of the SKA headquarters — a political and scientific prize — was due to be decided last week. It was a two-horse race: the United Kingdom or Italy. But the date came and went with no news. Astronomers have been left scratching their heads.

Nature has seen internal documents that explain both the delay and Cameron's expectancy. Italy won, and the United Kingdom kicked up a fuss. It threatened to pull out. It implied that Italy could not be relied on. It demanded (and will get) a recount. It acted, in other words, as a playground bully. Science has never been immune to the ugly reality of politics, but last week's unseemly gamesmanship is a particularly sorry example, and one that should not be allowed to stand unchallenged.

It is true that the Jodrell Bank Observatory near Manchester has acted as a temporary base for the SKA since 2012, and that the British would like that to continue. But the merits of two possible sites for a permanent home — the other is a historic observatory in Padua — have been the subject of an admirably transparent selection procedure, which the United Kingdom is now trying to undermine.

The two bids were assessed on a precise set of criteria, including political commitment to provide the extra financial support expected of a host and the quality of the research environment. The SKA board agreed on a timetable and chose a selection panel to assess the bids and recommend the winner. The panel comprised SKA board directors from three of the organization's 11 member countries — Australia, South Africa and the Netherlands — and a representative from the European Southern Observatory (ESO), an international astronomy organization headquartered in Garching, Germany. The panel's recommendation to the SKA board last week was crystal clear: both locations satisfied the criteria, but Padua was the better option.

When the United Kingdom saw that it had not won, it tried to change the rules, ramping up the pressure by circulating government-level letters to SKA board members. One from the head of the UK Science and Technology Facilities Council says "the [panel's] report does not appear to properly account for the scale and approval status of our financial commitment", and that "any decision on the headquarters must consider the broader ways this will affect the project and in particular the way in which it could affect the level of political commitment to the project". Another (unsigned) letter from the UK Department for Business Innovation and Skills says: "All things being equal — which they are in terms of meeting the HQ criteria — it makes no sense to dramatically increase the risk of the project by changing leadership from the UK to Italy … Transferring leadership would require the UK to radically re-assess participation in the project."

The SKA aims to use the globe as a giant radio telescope to image the early Universe, when the first stars and galaxies were forming. With a project this ambitious, it is perhaps not surprising that fights for control get dirty. In March 2012, South Africa was judged slightly better as a site to host the SKA telescope than Australia, but a political storm led to the decision to share the instruments. South Africa is building 3,000 dishes, and an even larger number of antennas are being installed in Australia.

Any competition for hosting the headquarters would have been undertaken in the knowledge that a non-UK winner would require a physical move. Italy may have a reputation among tabloid newspaper readers as Europe's clown — thanks in part to years under former prime minister Silvio Berlusconi — but hard-nosed scientists looking at its reliability in international scientific projects do not need to stoop to stereotypes. Italy is a reliable partner in both CERN, Europe's particle-physics lab near Geneva, Switzerland, and the ESO, for example. The country has competently headed organizations including the International Centre for Theoretical Physics and the International Centre for Genetic Engineering and Biotechnology for decades without problems.

"Science has never been immune to the ugly reality of politics."

Under pressure from the United Kingdom, the SKA board deferred a vote on the headquarters site to its next meeting at the end of April. It gave both countries until 20 March to submit extra material to the selection panel to confirm financial support, including their commitments if they are unsuccessful, and to address vague "operational and schedule matters; and organisational and reputational matters". The board also asked the panel to provide it with a comparative analysis "without an overall recommendation" by 10 April. These new criteria represent a move away from a transparent selection process to one that is based on murkier ground. ■

All in good time

Stratigraphers have yet to decide whether the Anthropocene is a new unit of geological time.

In western Berlin, Devil's Mountain rises 80 metres above the surrounding landscape to offer a clear view across the city. Known in German as *Teufelsberg*, the tree-covered hill looks primeval, but it was not there until 70 years ago. It was constructed as a dump for more than 25 million cubic metres of rubble cleared from the streets after the Second World War. So it is fitting that this artificial hill had a visit last year from a group of researchers assessing the geological imprint of humans on the planet.

The Anthropocene Working Group has a simple name but a very complicated job. These are the people who have to work out whether the world has entered a new slice of geological time — the Anthropocene.

As the group continues to assess the evidence, the rest of the planet has apparently made its decision. Three journals have been launched that are dedicated to research on the Anthropocene. Environmental advocates have heartily adopted the term and all it signifies, and so have many others, including artists and social scientists. And four years ago, *Nature* recommended that geologists formally accept the Anthropocene, arguing that the term "provides a powerful framework for considering global change and how to manage it" (see *Nature* **473,** 254; 2011).

But although many people have already made up their minds, those whose opinions matter the most have yet to do so (see pages 144 and 171).

The Anthropocene working group is diverse: about half of the three-dozen researchers are geologists, the rest a mix of archaeologists, palaeontologists, climate experts, atmospheric scientists and representatives of other disciplines. Working without pay over the past six years, and communicating mostly by e-mail, they have been sifting through evidence and arguments about when the Anthropocene might have begun, what kind of geological markers might define it, and whether it is worthy of recognition as a separate unit in Earth's geological history.

Despite the popular appeal of the Anthropocene, decisions relating to the geological timescale must rest with stratigraphers — researchers who study the evidence embedded in rock, ocean sediments, ice cores and other geological deposits. These people must look past the clamour and decide whether the Anthropocene is an appropriate new unit of chronostratigraphy. Their proposal will then be voted on by the International Commission on Stratigraphy (ICS) and the International Union of Geological Sciences.

The process remains conservative because the timescale is a tool used by tens of thousands of geoscientists around the world. Changes can create confusion, so the ICS requires strong scientific justification for any amendments. The fundamental question for the working group and for the ICS is whether geologists would find it sufficiently useful to define an Anthropocene unit in the rock record, which is the physical manifestation of the timescale. The Anthropocene would probably be an epoch that would sit after the Holocene, which started with the end of the last ice age, around 11,700 years ago.

If the Anthropocene is under way, then when did it start? Initial suggestions focused on the Industrial Revolution, but momentum has picked up to set the boundary after the Second World War. Since then, the global population has increased by 180%, water use by 215% and energy consumption by 375%. Researchers have called this surge the Great Acceleration, and it has skewed the composition of the atmosphere, warmed the planet, eroded the ozone layer and acidified the oceans. "The last 60 years have without doubt seen the most profound transformation of the human relationship with the natural world in the history of humankind," says the International Geosphere-Biosphere Programme, which has charted those changes.

"Stratigraphers must be given time and space to consider the consequences of formally adopting the Anthropocene."

It seems obvious that such broad planetary upheavals would warrant recognition on the geological timescale. But they may not be adequately reflected in stratigraphic evidence. In many parts of the globe, the geological record of the past 65 years is thin to non-existent. In the deep sea, less than a millimetre of sediment has built up, and that could be erased as ocean acidity increases. Signs of atmospheric changes are also preserved in recently laid down glacial ice, but much of that record could disappear in coming centuries as a result of global warming.

The working group still faces a considerable amount of work to evaluate whether — and how — to define the Anthropocene. If the committee or upper levels of the geology hierarchy decide against amending the timescale, the Anthropocene will not disappear. Many scientific disciplines and the public will continue to use the concept and word, in much the same way as they use the terms Neolithic era or Stone Age.

In the meantime, it is important that stratigraphers be given time and space to consider the consequences of formally adopting the Anthropocene. Any such change cannot be revisited for at least a decade, so the geological community will have to live with its decision for some time to come. ■

In the beginning

As the first true science journal marks 350 years, we must defend scholarly pursuits.

This month marks the 350th anniversary of arguably the first and longest-running scientific journal, *Philosophical Transactions: Giving Some Accompt of the Present Undertakings, Studies, and Labours of the Ingenious in Many Considerable Parts of the World*.

The first volume appeared on 6 March 1665, as a personal project of Henry Oldenburg, the first Secretary of the Royal Society in London, and was more of what many would regard as a magazine — with letters, book reviews and accounts of experiments from Europe's growing cadre of natural philosophers. Almost a century was to elapse before the Royal Society officially took it over and *Phil. Trans.* began to take its modern shape.

Part magazine and part journal, *Phil. Trans.* was much more than either. It was the journal — a genuinely new innovation — in which people of inquiring minds started to throw off the shackles of ancient received opinion and ask their own questions about the world around them. It was the start of scientific enquiry as we know it today.

By 1887, the breadth of scholarship had grown so much that *Phil. Trans.* could not encompass it all in one place. It split into streams — A and B — to cover separately the mathematical and physical sciences, and the biological sciences.

The schism was a sign of things to come. Today there are more than 40,000 scientific journals, from the hieratic to the demotic, the parochial to the cosmogonic. The arrival of electronic media is precipitating the biggest change in publishing since the invention of printing: journals are moving online, and access to knowledge, once the privilege of the educated European gentleman, is now increasingly seen as the right of any and every person — and rightly so. It would be all too easy to say that the only way now is onwards and upwards, as the bright light of enlightenment evaporates an ever-shrinking puddle of unreason.

Three and a half centuries of progress might seem a lot, but it is a tiny mote in the piebald passage of human history. Hard fought for, broad support for scholarly pursuit of a better world cannot be taken for granted.

The Library of Alexandria in Egypt was targeted and destroyed at various times between 48 BC and AD 642. For those inclined to dismiss such wanton vandalism as ancient history, think of the continuing and concerted efforts by many in the United States and elsewhere to sweep away science ranging from climate-change research to evolution. Consider that, as you read this, Islamist extremists are bulldozing the remains of ancient Assyria.

Even amid an almost uncountable profusion of journals, *Phil. Trans.* continues to thrive. All curious minds should wish it another 350 golden years. But the forces of irrationality are gaining in strength — one cannot afford to be complacent. ■

➲ NATURE.COM
To comment online, click on Editorials at:
go.nature.com/xhunqv

RICHARD A. ECKHARDT

Help to fight the battle for Earth in US schools

Scientists everywhere must champion a set of US education standards that promote Earth sciences, argues **Nicole D. LaDue**.

In another embarrassing moment for US scientists, Senator James Inhofe (Republican, Oklahoma) last month theatrically tossed a snowball onto the floor of the Senate during a debate on global warming. Despite all the talk of record temperatures in 2014, he said, there was snow on the lawn of the Capitol in Washington DC in winter.

Inhofe may or may not be aware of the distinction between weather and climate. Either way, he is unlikely to alter his views on climate change. More important is how such messages are received by the public, and in particular by the millions of schoolchildren who will be wrestling with the problem of global warming long after Inhofe is gone.

The United States has an opportunity to hugely improve the way that Earth sciences are taught in its schools. The difference between weather and climate, for example, could become standard discussion for third-grade classes, when children are eight or nine years old. Powerful lobby groups are trying to derail this opportunity. All scientists should help to stop them.

The quality of Earth-science education in most US schools is abysmal. I say this as a former high-school Earth-science teacher. Unlike physics, chemistry and biology, Earth science is typically taught by those with no adequate training in the subject.

In 2013, new standards were released that could reinvigorate US science education. Called the Next Generation Science Standards (NGSS), they were developed by scientists, science-education researchers and state-education representatives. In the NGSS, Earth science is on an equal footing with life science and physical science, from kindergarten through to the 12th grade (age 17 or 18). High-school students would learn how to "use a model to describe how variations in the flow of energy into and out of Earth's systems result in changes in climate". Imagine how well prepared the general public would be for making decisions about, and planning for, the impacts of climate change, nuclear waste disposal and investments in energy resources if they could "analyze geoscience data to make the claim that one change to Earth's surface can create feedbacks that cause changes to other Earth systems".

One truly exciting possibility about these standards relates to how they might be assessed. The testing movement has taken hold of US public education. Many state tests are predominantly multiple-choice, driving down the quality of classroom practice to memorization of facts and cookery-book laboratories. The NGSS will require new ways to assess both knowledge and scientific thinking. From a teacher's perspective, this provides an opportunity to teach science well and to engage students in the process of science, knowing that the assessments will challenge students to think rather than recall.

Imagine that the conversations in Washington DC moved beyond third-grade comprehension of daily weather versus average climate and focused on the complex economic impacts of climate change that we are already experiencing. This is possible if we put our efforts into adopting and implementing the NGSS appropriately across the country.

Under the US constitution, the federal government cannot tell states what to teach in schools. Each state must choose to adopt the NGSS, through approval by their boards of education or senate.

Currently, 13 states and Washington DC have adopted the NGSS; this covers about 14.5 million of the 50 million or so US students. However, state-level politics have blocked adoption in many cases. The National Center for Science Education, established to fight those who challenge the teaching of evolution and climate science across the United States, has been monitoring bills and lawsuits associated with the NGSS. In Kansas, there was a lawsuit over adoption of the standards because teaching evolution and the Big Bang was said to promote atheistic viewpoints. In Wyoming, Michigan and West Virginia, adoption has been challenged over the inclusion of anthropogenic climate change.

Even in states that have adopted the NGSS, hurdles remain. Many districts are looking to infuse the Earth-science content into physics, chemistry and biology classes, rather than establish high-quality Earth-science courses. This decision benefits the district because those classes prepare students for college-level courses, boosting national rankings. The teachers of physics, chemistry and biology are often unprepared for teaching the Earth-science content. If the NGSS is to succeed, science teachers must be trained on the content and develop or adjust the curriculum.

THE QUALITY OF EARTH-SCIENCE EDUCATION IN MOST US SCHOOLS IS ABYSMAL.

When scientists learn that my research focus is geoscience education, they lament the state of science literacy in the world around them. Certainly, if you are a US scientist, you probably feel that. But please do not tell me about your child's poorly prepared science teacher. Do not tell me that your undergraduate students are ill-prepared for college-level science.

Instead, tell me that you have asked your local school board how they are implementing the NGSS. Tell me that you have offered to run a workshop to teach your local teachers the Earth-science content they need. Tell me that your university department has written a letter to your state legislature on the importance of implementing Earth science in the NGSS to create a scientifically literate public prepared to make important decisions and pursue careers in high demand in your region. Stop complaining and do something. ■

➲ NATURE.COM
Discuss this article online at:
go.nature.com/qcpexr

Nicole D. LaDue *works in the Department of Geology and Environmental Geosciences at Northern Illinois University in DeKalb.*
e-mail: nladue@niu.edu

RESEARCH HIGHLIGHTS

CLIMATE CHANGE

Global warming could speed up

The rate of global warming could more than double over the coming decades, as greenhouse gases build up in Earth's atmosphere.

Steven Smith and his colleagues at the Pacific Northwest National Laboratory in College Park, Maryland, analysed the rate of warming in global climate simulations, and compared them over different 40-year periods. The team found that the global rate of warming in these simulations increases to an average of 0.25 °C per decade by 2020. An analysis of palaeoclimate data rarely showed rates of temperature change above 0.1 °C per decade during the last millennium.

The Arctic, Europe and North America will probably see larger increases in warming rates than the global average.

Nature Climate Change **http://dx.doi.org/10.1038/nclimate2552 (2015)**

ALEX HUIZINGA, NIS/MINDEN PICTURES/FLPA

ANIMAL BEHAVIOUR

Post-menopausal whales lead the hunt

After they reach menopause, female killer whales help their kin to survive by sharing their hunting expertise.

Humans, killer whales (*Orcinus orca*; **pictured**) and one other whale species are the only animals whose females are known to experience a long post-reproductive life. Female orcas can live into their 90s, even though they stop reproducing in their 40s. Darren Croft at the University of Exeter, UK, and his team analysed more than 750 hours of video footage of killer whales off the US Pacific coast collected between 2001 and 2009. Observations of 102 different whales up to 91 years old showed that post-reproductive females tended to lead group hunts for salmon, an important source of food. This leadership was particularly pronounced in years when salmon were scarce.

This is the first direct evidence that post-menopausal females are a source of ecological know-how, the authors say.

Curr. Biol. **http://doi.org/2mx (2015)**

PALAEONTOLOGY

Oldest *Homo* fossil found

A 2.8-million-year-old jawbone from Ethiopia may represent the earliest fossil from the genus *Homo* yet discovered — pushing back the known origins of humankind by nearly 500,000 years.

The fossil (**pictured**), analysed by Brian Villmoare at the University of Nevada, Las Vegas, William Kimbel at Arizona State University in Tempe and their colleagues, has key features of *Homo*, such as the parabolic shape of the jaw. But it also has more primitive traits, such as the jaw's overall size, that are seen in *Australopithecus afarensis*, a human ancestor that lived around 3 to 4 million years ago.

The fossil could belong to an ancestral *Homo* species, the authors say, filling a gap in the human fossil record.

Science **http://dx.doi.org/10.1126/science.aaa1343 (2015)**

ASTRONOMY

Quadruple images of supernova

A rare configuration of cosmic objects has produced multiple images of an exploding star in the same frame. If more images of the supernova appear, the system could provide a new way to measure the Universe's growth rate.

Patrick Kelly at the University of California, Berkeley, and his colleagues discovered the supernova kaleidoscope when examining images from the Hubble Space Telescope.

The multiple images occurred because two giant objects, a galaxy cluster and a galaxy within that cluster, acted as cosmic magnifying lenses that bent and boosted the light from the distant supernova. Light rays taking different paths around the gravitational lenses created the four different images. These rays took different amounts of time to travel their respective paths. Measuring such differences could help astronomers to better estimate

WILLIAM KIMBEL

distances in space and to measure the expansion of the Universe.
Science 347, **1123–1126 (2015)**

MICROBIOLOGY

Ultra small bacteria spotted

Bacteria roughly 1/100th the volume of a typical *Escherichia coli* have been found in groundwater.

Jillian Banfield at the University of California, Berkeley, Luis Comolli of Lawrence Berkeley National Laboratory in California, and their colleagues filtered groundwater through a mesh with holes around 0.2 micrometres in diameter and collected a variety of extremely small bacteria (around 0.009 cubic micrometres) that have never been cultured. Under the electron microscope, the microbes seemed to have tightly packed DNA, few of the protein-making structures called ribosomes, and structures that might allow the cells to connect and communicate with one another.

The researchers suggest that these bacteria had not been cultured before because they depend on other microbes to grow.
Nature Commun. 6, **6372 (2015)**

PHOTONIC MATERIALS

Pulled fibres shift colour

Rubbery fibres have been developed that reversibly change colour when stretched or bent.

Xuemei Sun, Huisheng Peng and their collaborators at Fudan University in Shanghai, China, attached microscopic plastic spheres to elastic fibres that were wound with carbon nanotubes. As the fibre stretches, the spaces between the microspheres increase in size along the length of the fibre, whereas they decrease in the radial direction. This changes the wavelengths of light that are reflected by the fibres, resulting in shifts in colour between red, green and blue as the fibre is stretched and released. The fibres remained stable after 1,000 rounds of stretching and were woven into fabric in various patterns.

Such 'mechanochromic' materials could be used in wearable displays or sensors, the authors say.
Angewandte Chemie **http://doi.org/f259np (2015)**

GEOLOGY

Hydration lifts Earth's crust

The high elevation of parts of the western United States could be a result of water percolating up from deep in Earth's crust, and changing the crust's mineral composition, making the rocks more buoyant.

Geologists have been hard-pressed to explain why Colorado and much of Wyoming have lifted by more than 2 kilometres over the past 75 million years. A team led by Craig Jones at the University of Colorado Boulder reanalysed data on the geology and seismology of the region and conclude that in lower regions, such as Montana, fragments of crustal rock contain dense minerals such as garnet. Beneath high-elevation areas, however, the rocks contain a different suite of less dense minerals. The authors suggest that these were produced by water reacting with the dense minerals and so making the crust lighter.

The water may have come from the dehydration of a deeply buried, ancient crustal slab.
Geology **http://doi.org/2ps (2015)**

STRUCTURAL BIOLOGY

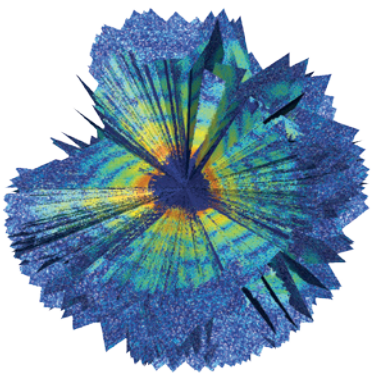
X-rays reveal virus innards

With the help of powerful X-rays, researchers have determined the three-dimensional structure of a single giant virus particle. This shows how tiny objects that cannot be easily crystallized can still be imaged in 3D.

X-ray crystallography is commonly used to work out the structure of molecules, but these must be crystallized first. However, free-electron lasers generate such high-energy X-ray pulses that they can, in theory, produce pictures of just a single molecule.

Tomas Ekeberg at Uppsala University in Sweden and his colleagues fired these lasers at single particles of the *Acanthamoeba polyphaga* mimivirus. They used algorithms to combine X-ray diffraction patterns from many specimens and created a 125-nanometre-resolution image of the virus (**pictured**).

The results confirm that that the mimivirus is less densely packed with genetic material than smaller viruses tend to be.
Phys. Rev. Lett. 114, **098102 (2015)**



TOMAS EKEBERG, UPPSALA UNIV.

SOCIAL SELECTION

Popular articles on social media

Scientific art kicks off Twitter storm

Images of painted pterosaurs, ceramic diatoms and quilts depicting neurons flooded scientists' Twitter feeds, after the writers of Symbiartic, *Scientific American*'s art blog, launched SciArt Week on 1 March. Researchers and artists posted a flurry of artwork highlighting the beautiful side of science, using the hashtag #sciart.

Malcolm Campbell, a plant scientist at the University of Toronto, Scarborough, Canada, was one of the first researchers to announce SciArt week on Twitter. "Art captures the imagination in a way that science alone cannot," he says. "It's a wonderful way to make science more tangible to the public."

NATURE.COM
For more on popular papers:
go.nature.com/r9apn

CLIMATE-CHANGE BIOLOGY

Insects feast under high CO_2

Leaf-eating insects in northern temperate forests consume more of the forest canopy when carbon dioxide levels are increased, which could limit forests' capacity to act as carbon sinks in a warming world.

John Couture and his colleagues at the University of Wisconsin-Madison, found that in parts of a research forest exposed to raised CO_2 levels, herbivorous insects increased their consumption of foliage by 88%. This led to an average of 70 grams of carbon-sequestering biomass lost per square metre of forest per year.

Increased CO_2 could be causing this effect by changing the nutrient content of leaves and also by boosting the number of leaf-eating insects, the authors say. They also suggest that insect behaviour should be incorporated into models that estimate the effects of high CO_2 on forest productivity.
Nature Plants **http://dx.doi.org/10.1038/nplants.2015.16 (2015)**

NATURE.COM
For the latest research published by *Nature* visit:
www.nature.com/latestresearch

SEVEN DAYS *The news in brief*

EVENTS

Dawn arrival

NASA's Dawn spacecraft slipped into the gravitational pull of Ceres on 6 March, making it the first probe to visit a dwarf planet. At nearly 1,000 kilometres across, Ceres, located in the asteroid belt, is one of the largest unexplored worlds in the Solar System. Dawn will orbit Ceres for the next 15 months, gathering information about the large amounts of water thought to lurk within the asteroid. The craft also visited the asteroid Vesta in 2011–12; its arrival at Ceres also makes Dawn the first probe to have orbited two celestial bodies. See go.nature.com/uwg9fb for more.

Animal research

More than 120 research institutes, organizations and societies in Europe called on the European Commission on 4 March to oppose an initiative calling for a complete ban on research using animals. Animal-rights activists submitted a petition to the commission on 3 March signed by more than 1.1 million citizens. As part of a European Citizens' Initiative, the petition opens a procedure for a hearing in the European Parliament, and for reconsideration of legislation. In a joint statement, the bodies supported the current legislation, saying that it guarantees high standards of animal welfare while allowing crucial health research.

Tardy, weak El Niño

A weak El Niño pattern has developed several months later than normal in the equatorial Pacific Ocean, forecasters with the US National Oceanic and Atmospheric Administration (NOAA) announced on 5 March. Marked by warmer than average waters, El Niño conditions can have far-flung consequences, from greater precipitation in the southeastern United States to droughts in southeast Asia. NOAA says that there is a 50–60% chance that El Niño conditions will continue into the Northern Hemisphere summer, but that the system is too weak and too late to have major global impacts. See go.nature.com/qbmdci for more.

CARL DE SOUZA/AFP/GETTY

Ivory stockpile burns in Kenya

Fifteen tonnes of ivory were burned in Nairobi National Park on 3 March, as Kenya became the latest country to destroy seized stocks to deter elephant poachers. At the burn, President Uhuru Kenyatta said that the country would soon destroy the rest of its stockpile, too. The country's previous president burned around 5 tonnes of ivory in 2011. China said last month that it would ban all imports of ivory, as poaching continues to kill hundreds of elephants in Africa every week.

PEOPLE

ITER head

Bernard Bigot was appointed director-general of ITER, a project to build the world's biggest nuclear-fusion reactor in southern France, at a meeting in Paris on 5 March. Bigot, who retired as chairman of the French Alternative Energies and Atomic Energy Commission (CEA) in January, was nominated for the ITER post last November (see *Nature* http://doi.org/2q3; 2014). He has promised reforms of ITER's complex multinational management, and to address the project's schedule slippages and cost increases. Bigot begins his five-year term immediately.

Cancer chief

The director of the US National Cancer Institute (NCI), Harold Varmus, announced on 4 March that he will step down after five years in the post. Varmus will leave the centre, part of the National Institutes of Health (NIH), at the end of the month, and plans to open a lab at the Weill Cornell Medical College in New York City. He was director of the NIH from 1993 to 1999, and won the 1989 Nobel Prize in Physiology or Medicine for his work on the role of retroviruses in cancer. NCI deputy director Douglas Lowy will serve as interim chief until a replacement is appointed. See go.nature.com/sv4ful for more.

BUSINESS

Cancer-drug firm

Pharmaceutical firm AbbVie agreed to pay US$21 billion to purchase Pharmacyclics, a company that specializes in cancer drugs, in a deal announced on 4 March. Pharmacyclics, based in Sunnyvale, California, makes Imbruvica (ibrutinib), a blood-cancer drug that targets

a protein called Bruton's tyrosine kinase, and which brought in $548 million in 2014. AbbVie, of North Chicago, Illinois, plans to close the deal in the middle of 2015.

'Biosimilar' drug

The US Food and Drug Administration awarded its first approval to a 'biosimilar' drug on 6 March. The drug, Zarxio (filgrastim-sndz), is similar to a previously approved protein used to prevent infections following cancer chemotherapy. The Zarxio decision could herald the approval of other biosimilars, and reduce health-care costs. Zarxio, made by the generics arm of the Swiss pharmaceutical firm Novartis, was approved in Europe in 2009, but the United States has struggled to formulate regulations governing biosimilars. See go.nature.com/omxrup for more.

TECHNOLOGY

Solar plane

Swiss pilots launched an attempt on 9 March to fly around the world in a plane powered only by solar energy. Bertrand Piccard and André Borschberg began their trip in the experimental plane Solar Impulse 2 (**pictured**) in Abu Dhabi. The plane, which has a wingspan wider than

JEAN REVILLARD VIA GETTY

that of a jumbo jet but is the weight of a small car, uses more than 17,000 solar cells and rechargeable lithium-ion batteries to fly for several days and nights in a row. The five-month trip will include passing over both the Atlantic and Pacific oceans.

FUNDING

Australian crisis

Much of Australia's shared national research infrastructure is under threat of closure because of uncertainty over whether it will receive the Aus$150 million (US$116 million) allocated by the government last year. Organizations representing Australian scientists wrote an open letter to Australia's Prime Minister Tony Abbott on 4 March warning of the crisis. Twenty-seven facilities under the National Collaborative Research Infrastructure Strategy, employing some 1,700 research staff, will close if funding does not come through. The cash is tied to controversial legislation on higher-education reform that has not yet passed through parliament. See go.nature.com/3q8eiq for more.

RESEARCH

Brain project

The European Commission has recommended changes to the governance of Europe's €1-billion (US$1.1-billion) Human Brain Project, which brings together neuroscience and computing. A summary report published by the commission on 6 March states that the decision-making processes need to be made "simple, fair and transparent". Similar recommendations were made on 9 March by an independent mediation committee that was analysing deep rifts in the project. See go.nature.com/knoaqq for more.

EU funding scrutiny

Bulgaria has agreed to have its deficient research system scrutinized by a group of international science-policy experts on behalf of the European Commission. The review, scheduled to begin in April, will be the first carried out under the auspices of the commission's Policy Support Facility, a €20-million (US$22-million) programme launched on 3 March with the goal of strengthening science and innovation capacities in the European Union.

COMING UP

14–18 MARCH
The decadal UN World Conference on Disaster Risk Reduction will take place in Sendai, Japan. It aims to help countries prepare for disasters.
go.nature.com/opisic

15 MARCH
NASA's Super Pressure Balloon is scheduled to launch after this date from Wanaka, New Zealand. The research balloon aims to break the previous record of 54 days in flight, and will also test technology developed by NASA over 15 years.

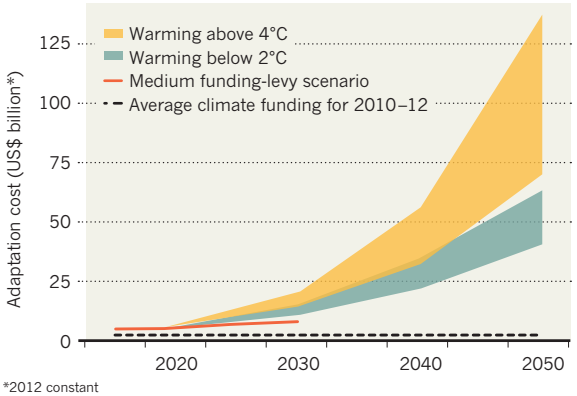
17–21 MARCH
The UN World Conference on Tobacco or Health takes place in Abu Dhabi. The event, which occurs every three years, will focus on the link between tobacco use and non-communicable diseases that kill 38 million people each year.
go.nature.com/smiiek

TREND WATCH

The costs to Africa of adapting to climate change could rise to between US$50 billion and $100 billion per year by 2050, depending on global efforts to reduce greenhouse-gas emissions, the United Nations Environment Programme reported on 4 March. It estimates that current annual financial aid is just $1 billion to $2 billion. Levies on sectors such as tourism and banking could raise $4.8 billion per year, but even if current policies keep warming to below 2 °C, costs could still outpace revenue as early as 2020.

CLIMATE-CHANGE COSTS

Adapting to the effects of climate change will cost Africa dearly, even if the world acts on warming.



*2012 constant

SOURCE: UNEP

NATURE.COM
For daily news updates see:
www.nature.com/news

NEWS IN FOCUS

CHINAFOTOPRESS/GETTY

Health workers swab a pigeon in a market in Changsha, China.

VIROLOGY

Flu genomes trace H7N9 evolution

But surveillance of avian influenza viruses is patchy and slow.

BY DECLAN BUTLER

No one knows the pandemic potential of the H7N9 avian influenza that has infected more than 560 people in China and killed 204 since it was first detected in March 2013. But the largest-ever genomic survey of the virus in poultry now provides a more detailed picture of its evolution and spread.

Such information can help to target control efforts, and to monitor the evolution of the virus. But an analysis of sequences submitted to the GenBank repository in the past 15 years suggests that genetic surveillance of avian flu viruses in birds is patchy and less than prompt.

For now, H7N9 flu does not spread easily among people. But as with many bird-flu viruses, a concern is that it could evolve to do so.

In a paper published on *Nature*'s website this week (T. T.-Y. Lam *et al*. *Nature* http://dx.doi.org/10.1038/nature14348; 2015), an international team of researchers describes how it tracked the virus from October 2013 to July 2014 by taking swabs from poultry at live-bird markets in 15 cities over 5 provinces in eastern China. The group detected the virus in markets in seven cities and in 3% of samples on average.

The team then sequenced the genomes of 438 viral isolates and found that as the virus spread south, it evolved into three main branches, with multiple sub-branches.

Such diversification is expected, but tracking it can help to identify the main trade routes and markets that fuel a virus's spread. "The extent of viral transmission among chickens was largely unclear until our paper showed that the virus had diverged into regional lineages," says Yi Guan, a co-author of the paper and a virologist at the State Key Laboratory of Emerging Infectious Diseases in Shenzhen, China. "Eastern China remains as a reservoir and 'distribution centre' for this virus," he says.

Despite such insights, relatively few sequences of H7N9 have been collected. Sequences from only eight H7N9 viral isolates collected from birds in 2014 have been deposited in GenBank. That is not enough for geographical mapping of the virus over time, says Marius Gilbert, an avian-flu epidemiologist and ecologist at the French-speaking Free University of Brussels.

Nor is the latest paper up to date. A new winter wave of H7N9 is under way, and probably has different patterns of spread.

In 2012, *Nature*'s news team reported that genetic surveillance of animal-flu viruses is patchy globally: most genomes are sequenced months or years after collection (see *Nature* **483,** 520–522; 2012). Current GenBank data suggest that this is still true. Far more flu sequences are being deposited in GenBank, but many are from samples collected some time ago.

Guan agrees that timely monitoring is important. But surveillance and viral sequencing are costly and time-consuming, and for H7N9 require access to a biosafety-level-3 lab. Given the complications, Guan thinks that the number of recent H7N9 sequences is not grossly low.

Adding to the time lag, public authorities and researchers who sequence flu strains sometimes make the data public only when, or if, they publish — so sequences can languish. The authors of the latest study have sent sequences to GenBank and had already shared the data with the World Health Organization and other bodies.

Guan and his co-authors warn that H7N9 "should be considered as a major candidate to emerge as a pandemic strain". But predicting pandemic potential is an embryonic science. Last year, a prominent international group of researchers argued that there is little evidence that flu viruses that cause sporadic human infections are a greater pandemic threat than viruses that have not yet infected humans (C. A. Russell *et al*. *eLife* **3,** e03883; 2014). But Guan says that given the vast number of flu viruses, it is necessary to prioritize targets for control and vaccine development — and that H7N9 should be high on that list. ■

POLICY

Mistrust and meddling unsettles US science agency

National Science Foundation under pressure from lawmakers to revise its agenda.

BY BOER DENG

GREG E. MATHIESON SR/REX

National Science Foundation head France Córdova (left) is trying to improve lawmakers' view of the agency.

The US National Science Foundation (NSF) has had a tough couple of years. Republicans in the US Congress have put the agency under the microscope, questioning its decisions on individual grants and the purpose of entire fields of study. The agency was without a permanent director for a year, and it is now planning an expensive, and controversial, move to new headquarters.

As she prepares to mark one year at the agency's helm, astrophysicist France Córdova is carefully navigating these challenges. "I used to be a mountaineer," she says. "It's all about looking at every move and how you can best do it so that you don't take a fall." But many researchers worry that Congress has begun to interfere with the scientific process. As mistrust grows, the NSF is caught between the scientists it serves and the lawmakers it answers to.

Córdova has moved aggressively to repair relations with Congress. Aides to lawmakers who participated in a December trip to NSF facilities in Antarctica say that the journey was successful. And to address concerns about transparency, the agency has instituted guidelines that should make its grant summaries easier to understand.

But such efforts seem to have had little influence on an investigation of the NSF's funding decisions by Representative Lamar Smith (Republican, Texas), chairman of the House Committee on Science, Space, and Technology. Since he took the job two years ago, Smith has sought to root out what he sees as wasteful spending by the US$7-billion NSF. He has introduced legislation that would require the agency to certify that every grant it awards is in the "national interest", and he has repeatedly sought, and been given, confidential information about individual NSF grants — albeit in redacted form. On at least four occasions, staff from the science committee travelled to the NSF's headquarters in Arlington, Virginia, to review such documents, most recently on 28 January.

"There is a sense of exhaustion among researchers as this has continued," says Meghan McCabe, a legislative-affairs analyst at the Federation of American Societies for Experimental Biology in Bethesda, Maryland. An NSF programme director who asked not to be named is more direct: "Having them in our building questioning our work like that felt like an attack."

But Córdova argues that the political landscape has changed and the NSF must adapt. "Congress absolutely has the right to request whatever materials for oversight they want," she says. "Just because we're not used to it doesn't mean it's a violation." At a House subcommittee hearing in February, Córdova told lawmakers that she supports Smith's proposal to require that NSF grants support the national interest. (The NSF already judges grant applications on their potential "broader impacts" as well as on scientific merit; in December, it began asking applicants to articulate how their projects serve the national interest, as defined by the agency's mission statement.)

Among scientists, however, there is anxiety that Córdova has been too conciliatory towards critics in Congress. "Once you start to compromise, you're just inviting harassment," says Lloyd Etheredge, a social scientist at the Policy Sciences Center in Bethesda.

Some argue that the concessions to Congress will compromise the agency's peer-review process. "If the NSF is funding a grant, it should by definition be in the national interest," says John Bruer, president emeritus of the James S. McDonnell Foundation in St Louis, Missouri, who in 2011 led an NSF task force on grant criteria. "When you add stuff about the national interest, you are potentially inviting criteria apart from judging the best science."

MANY PRIORITIES

The agency's ongoing struggle with Congress has left Córdova with less time to deal with internal challenges, such as employees who are disgruntled by a 2013 decision to move NSF's headquarters from a suburb close to Washington DC to a site that is farther away and has smaller facilities for some staff. In October, a federal government arbitrator sided with an NSF employee union and ordered the agency to revise its design to accommodate large, individual workspaces in the new headquarters. Córdova has sought to address unrest about the move through a series of meetings and working

groups, but rumours persist that many senior employees will opt to retire rather than relocate.

That would be a significant blow to an agency that is already stretched. The NSF's budget has grown slowly but steadily in recent years, reaching $7.3 billion in fiscal year 2015. But even though the number of grant proposals submitted to the agency has risen by 65% over the past 15 years, the NSF has seen only a 20% increase in the number of full-time employees.

The resulting increase in workload has affected staff morale. A 2014 survey by the US Office of Personnel Management found that only 45% of NSF employees felt that the agency's leadership generated "high levels of motivation and commitment in the workforce", compared with 53% in 2010. And just over one-third of workers were negative about the opportunities available for getting a better job at the agency.

As Córdova enters the second year of her six-year term, the challenges ahead are clear.

Eugene Skolnikoff, a political scientist at the Massachusetts Institute of Technology in Cambridge, says that winning and maintaining the trust of the scientific community gives an NSF director clear authority to negotiate with Congress. "The best NSF directors," he says, "have been the ones who really got the staff and the scientists behind their vision." ■

BIOLOGY

DNA clock proves tough to set

Geneticists meet to work out why the rate of mutation in the human genome is hard to pin down.

BY EWEN CALLAWAY

Mathematicians keep refining π even though they know it to more than 12 trillion digits; physicists beat themselves up because they cannot pin down the gravitational constant beyond three significant figures. Geneticists, by contrast, are having trouble deciding between one measure of how fast human DNA mutates and another that is half that rate.

The rate is key to calibrating the 'molecular clock' that puts DNA-based dates on events in evolutionary history. So at an intimate meeting in Leipzig, Germany, on 25–27 February, a dozen speakers puzzled over why calculations of the rate at which sequence changes pop up in human DNA have been so much lower in recent years than previously. They also pondered why the rate seems to fluctuate over time. The meeting drew not only evolutionary geneticists, but also researchers with an interest in cancer and reproductive biology — fields in which mutations have a central role.

"Mutation is ultimately the source of all heritable diseases and all biological adaptations, so understanding the rate at which mutations evolve is a fundamental question," says Molly Przeworski, a population geneticist at Columbia University in New York City who attended the Human Mutation Rate Meeting.

Researchers tried to put a number on the human mutation rate even before they knew that genetic information is encoded in DNA. In the 1930s, pioneering geneticist J. B. S. Haldane came up with a good estimate by measuring how the mutations responsible for haemophilia appeared in extended families.

Later estimates of the mutation rate counted the differences between stretches of DNA and protein amino-acid sequences in humans and those in chimpanzees or other apes, and then divided the number of differences by the time that has elapsed since the species' most recent common ancestor appeared in the fossil record. These estimates were clouded by the patchiness of the fossil record, but researchers eventually settled on a consensus: each DNA letter, on average, mutates once every billion years. That is a "suspiciously round number", molecular anthropologist Linda Vigilant of the Max Planck Institute for Evolutionary Anthropology in Leipzig told *Nature* in 2012 (see *Nature* **489,** 343–344; 2012).

"The fact that the clock is so uncertain is very problematic for us."

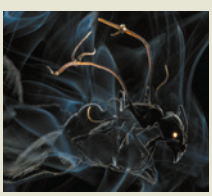
In the past six years, more-direct measurements using 'next-generation' DNA sequencing have come up with quite different estimates. A number of studies have compared entire genomes of parents and their children — and calculated a mutation rate that consistently comes to about half that of the last-common-ancestor method.

A slower molecular clock worked well to harmonize genetic and archaeological estimates for dates of key events in human evolution, such as migrations out of Africa and around the rest of the world[1]. But calculations using the slow clock gave nonsensical results when extended further back in time — positing, for example, that the most recent common ancestor of apes and monkeys could have encountered dinosaurs. Reluctant to abandon the older numbers completely, many researchers have started hedging their bets in papers, presenting multiple dates for evolutionary events depending on whether mutation is assumed to be fast, slow or somewhere in between.

Last year, population geneticist David Reich of Harvard Medical School in Boston, Massachusetts, and his colleagues compared the genome of a 45,000-year-old human from Siberia with genomes of modern humans and came up with the lower mutation rate[2]. Yet just before the Leipzig meeting, which Reich co-organized with Kay Prüfer of the Max Planck Institute for Evolutionary Anthropology, his team published a preprint article[3] that calculated an intermediate mutation rate by looking at differences between paired stretches of chromosomes in modern individuals (which, like two separate individuals' DNA, must ultimately trace back to a common ancestor). Reich is at a loss to explain the discrepancy. "The fact that the clock is so uncertain is very problematic for us," he says. "It means that the

MORE ONLINE

PICTURES OF THE MONTH



Zombie ants, mummified seals and a gorilla punch go.nature.com/mn4fsx

MORE NEWS

- No link between psychedelics and psychosis go.nature.com/cwramz
- How Ebola survivors mustered an immune defence go.nature.com/6d3upn
- Complex societies evolved without belief in all-powerful deity go.nature.com/kumjxo

STORY OF THE WEEK

Ethiopian jawbone may mark dawn of humankind go.nature.com/icyibm

ANAND VARMA/NATIONAL GEOGRAPHIC/WORLD PRESS PHOTO

KAYE REED/SCIENCE/AAAS

dates we get out of genetics are really quite embarrassingly bad and uncertain."

Reich hoped that even if the meeting did not reach a consensus on mutation rate, it would highlight the research that is needed to move forward. He and Prüfer kicked off the meeting by polling attendees on their favoured rate, and found that the lower figure had gained popularity, but there was still a wide spread of opinions.

Increasingly, Reich and others conclude that the human mutation rate has fluctuated over millions of years. Much of the discussion at the meeting revolved around when it accelerated and decelerated — and why. Evolutionary changes in metabolism or reproductive biology are both possible causes. Aylwyn Scally, a population geneticist at the University of Cambridge, UK, thinks that the common ancestor of great apes, which lived between 20 million and 12 million years ago, had longer generations than its relatives on the monkey branch of the primate family tree. That would have slowed mutation: a longer generation would lead to fewer mutations per year, on average.

Medical-minded geneticists also fret about mutation rates. Meeting attendee Michael Stratton, director of the Wellcome Trust Sanger Institute in Hinxton, UK, is a cancer geneticist who studies the causes of DNA mutations. Environmental agents such as tobacco smoke trigger some cancers, but others are caused by the normal biochemical operations of cells — through processes that are little-known, says Stratton. Working out what these are could explain fluctuations in the mutation rate.

Reproductive biologists are also interested in the human mutation rate — in part because they have found that some diseases are more common in the children of older men than of younger ones. Sperm are produced throughout a man's life, whereas women are born with a full array of eggs. The constant division of sperm precursor cells means that men tend to pass on more new mutations to their offspring than women — four times as many, according to a 2012 estimate[4] — and older fathers transmit more mutations than young ones. This means that changes in the biology of sperm production or paternal age over evolutionary time could influence mutation rate.

Even though the human mutation rate is still uncertain and unstable, Reich proposed at the meeting that researchers use the slower value for their work, at least until better data come along. Just don't think of it as a constant, he cautions: "This is not the speed of light. This is not physics." ■

1. Scally, A. & Durbin, R. *Nature Rev. Genet.* **13,** 745–753 (2012).
2. Fu, Q. *et al*. *Nature* **514,** 445–449 (2014).
3. Lipson. M. *et al.* Preprint at http://dx.doi.org/10.1101/015560 (2015).
4. Kong, A. *et al*. *Nature* **488,** 471–475 (2012).

FRANK HURLEY/SCOTT POLAR RES. INST./UNIV. CAMBRIDGE/GETTY

The Grytviken whaling station on South Georgia island in the First World War. It has long been abandoned.

MARINE ECOLOGY

World's whaling slaughter tallied

Commercial hunting wiped out almost three million animals last century.

BY DANIEL CRESSEY

The first global estimate of the number of whales killed by industrial harvesting last century reveals that nearly 3 million cetaceans were wiped out in what may have been the largest cull of any animal — in terms of total biomass — in human history.

The devastation wrought on whales by twentieth-century hunting is well documented. By some estimates, sperm whales have been driven down to one-third of their pre-whaling population, and blue whales have been depleted by up to 90%. Although some populations, such as minke whales, have largely recovered, others — including the North Atlantic right whale and the Antarctic blue whale — now hover on the brink of extinction.

But researchers had hesitated to put a number on the global scale of the slaughter. That was largely because they did not trust some of the information in the databases of the International Whaling Commission, the body that compiles countries' catches and that manages whaling and whale conservation, says Robert Rocha, director of science at the New Bedford Whaling Museum in Massachusetts.

Rocha, together with fellow researchers Phillip Clapham and Yulia Ivashchenko of the National Marine Fisheries Service in Seattle, Washington, has now done the maths, in a paper published last week in *Marine Fisheries Review* (R. C. Rocha Jr, P. J. Clapham and Y. V. Ivashchenko *Mar. Fish. Rev.* **76,** 37–48; 2014). "When we started adding it all up, it was astonishing," Rocha says.

The researchers estimate that, between 1900 and 1999, 2.9 million whales were killed by the whaling industry: 276,442 in the North Atlantic, 563,696 in the North Pacific and 2,053,956 in the Southern Hemisphere. Other famous examples of animal hunting may have killed greater numbers of creatures — such as hunting in North America that devastated bison and wiped out passenger pigeons. But in terms of sheer biomass, twentieth-century whaling beat them all, Rocha estimates.

"The total number of whales we killed is a really important number. It does make a difference to what we do now: it tells us the number of whales the oceans might be able to support," says Stephen Palumbi, a marine ecologist at Stanford University in California. He thinks that 2.9 million whale deaths is a "believable" figure.

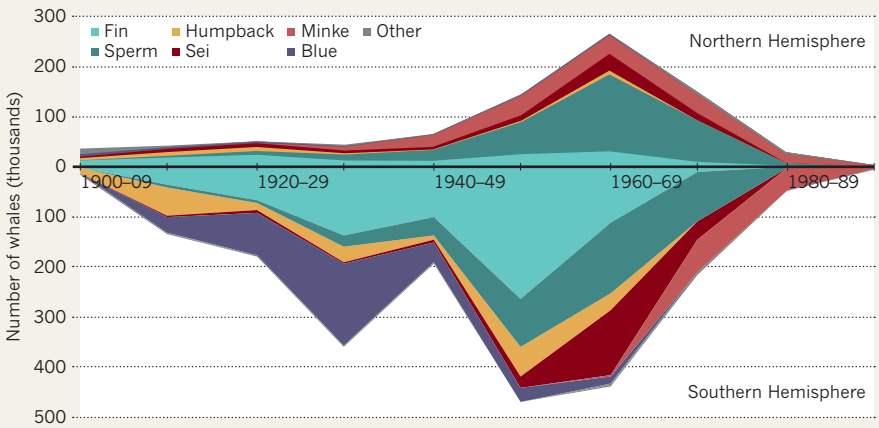
Sail-powered whaling ships took around 300,000 sperm whales between the early 1700s and the end of the 1800s. But with the aid of diesel engines and exploding harpoons, twentieth-century whalers matched the previous two centuries of sperm-whale destruction in just over 60 years. The same number again were harvested in the following decade. As one whale species became depleted, whalers would switch to another (see 'The largest hunt'). Most commercial hunting was put on hold only in the 1980s.

"It's an eye-opener for people to understand just how many whales were killed in the twentieth century alone. It shows how methodical and efficient whalers were," says Howard Rosenbaum, a cetacean researcher who runs the Ocean Giants Program at the Wildlife Conservation Society, a non-governmental organization headquartered in New York City.

The latest estimate depended on detective work by Ivashchenko, who documented a huge illegal whaling operation in the Northern Hemisphere by the former Soviet Union for her 2013 doctoral thesis. Through interviews with former Soviet whalers and researchers, and reports from the whaling industry that she uncovered, she found that more than half a million whales had been caught by Soviet vessels, and that 178,811 of those were never declared to the International Whaling Commission.

THE LARGEST HUNT

Industrial whaling vessels killed nearly 2.9 million animals of various species in the twentieth century. Most were fin and sperm whales, but blue, sei, humpback and minke whales were also taken in their thousands.



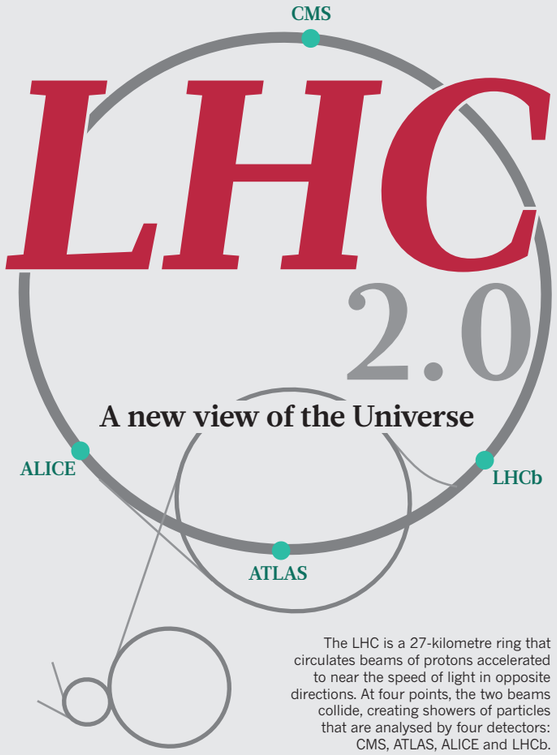
SOURCE: *MAR. FISH. REV.* **76**, 37–48 (2014)

Some researchers have used genetic data on certain populations to estimate how many whales existed before human hunting began. But the genetics has often suggested much larger original populations than the whaling records imply, says Rosenbaum. The estimates are now creeping closer together, he adds, as the genetics work improves and the catch data are revised upwards with inclusion of the true Soviet figures and other revisions. Understanding how many whales were taken from the oceans might mean that targets that define when a species has recovered need to be changed, he says.

Rocha adds that 2.9 million whales is a lower bound. Although motorized boats were more efficient than the original sailing vessels in capturing whales, some of the animals they mortally wounded would escape or not make it onto official records. "The actual number of whales killed is going to be more," he says. ■

LHC 2.0

A new view of the Universe



The LHC is a 27-kilometre ring that circulates beams of protons accelerated to near the speed of light in opposite directions. At four points, the two beams collide, creating showers of particles that are analysed by four detectors: CMS, ATLAS, ALICE and LHCb.

The world's most powerful particle collider is poised to roar once again into action after a two-year hiatus. At the end of March, the Large Hadron Collider (LHC) at CERN, Europe's particle-physics lab near Geneva, Switzerland, will start smashing particles together at a faster rate and with higher energies than ever before. "We're standing on the threshold of a completely new view of the Universe," says Tara Shears, a particle physicist at the University of Liverpool, UK.

The first run began in earnest in November 2009 and ended in February 2013. The LHC collided particles — mainly protons but also heavier particles such as lead ions — at high enough energies to discover the Higgs boson in 2012, which garnered those who predicted the subatomic particle a Nobel prize.

In the next run, set to last three years, energies will rise to an eventual 14 teraelectronvolts (TeV; see 'Hardware rebooted'). One hope is that higher energies will produce evidence for supersymmetry, an elegant theory that could extend the standard model of particle physics (see 'Desperately seeking SUSY'). They could also shake out particles of dark matter, the invisible substance that is thought to make up 85% of the matter in the Universe (see 'Decays decoded').

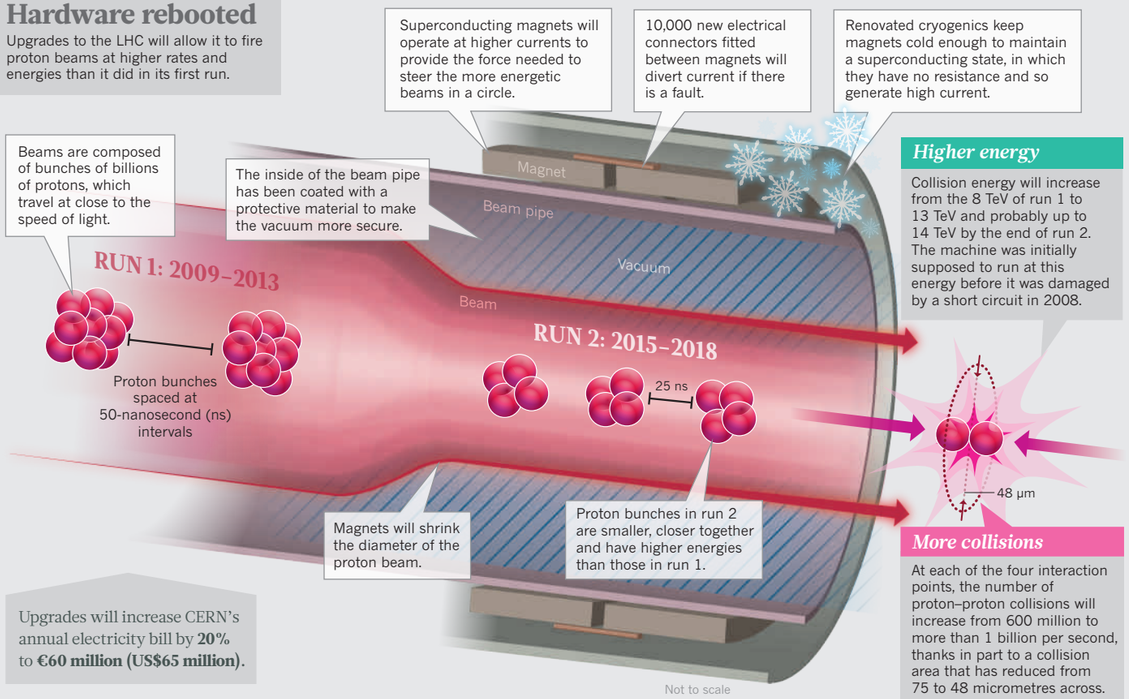
More collisions will enable more-precise study of the Higgs' nature (see 'The Higgs factory') and will provide clarity on anomalies hinted at in run 1 (see 'Known unknowns').

"In the first run we had a very strong theoretical steer to look for the Higgs boson," says Shears. "This time we don't have any signposts that are quite so clear."

BY ELIZABETH GIBNEY / ILLUSTRATION BY NIK SPENCER

Hardware rebooted

Upgrades to the LHC will allow it to fire proton beams at higher rates and energies than it did in its first run.



Higher energy

Collision energy will increase from the 8 TeV of run 1 to 13 TeV and probably up to 14 TeV by the end of run 2. The machine was initially supposed to run at this energy before it was damaged by a short circuit in 2008.

More collisions

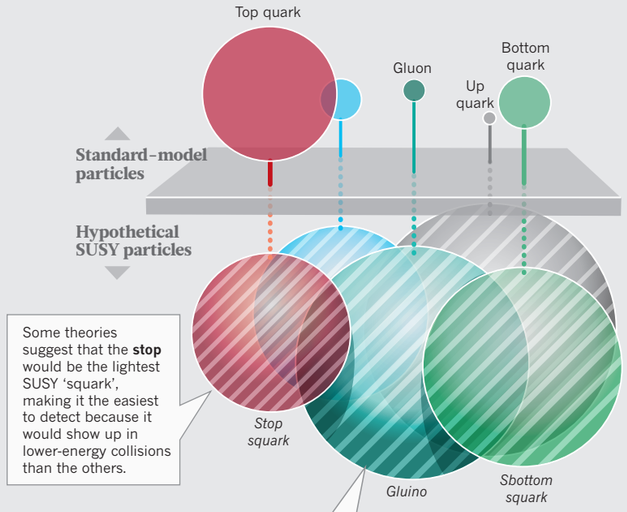
At each of the four interaction points, the number of proton–proton collisions will increase from 600 million to more than 1 billion per second, thanks in part to a collision area that has reduced from 75 to 48 micrometres across.

Upgrades will increase CERN's annual electricity bill by **20%** to **€60 million (US$65 million)**.

Higher energy

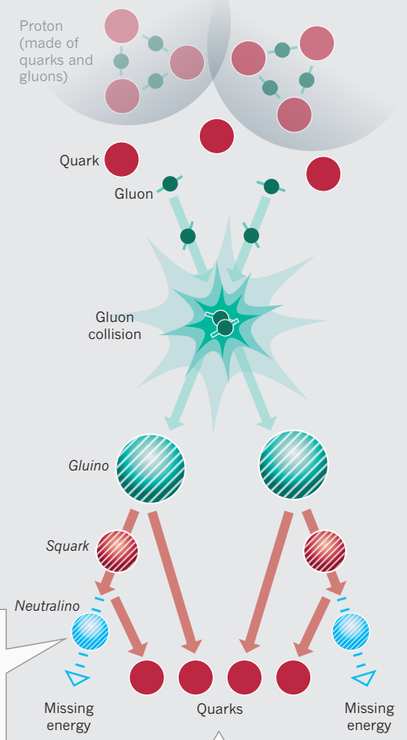
Desperately seeking SUSY

Higher energies mean that the LHC can produce heavier particles (because of $E = mc^2$) — and perhaps some of those predicted by the theory of supersymmetry, or SUSY. An extension to the standard model of particle physics, SUSY postulates a giant 'superpartner' for each known particle, and would offer explanations for mysteries such as the nature of dark matter.



Decays decoded

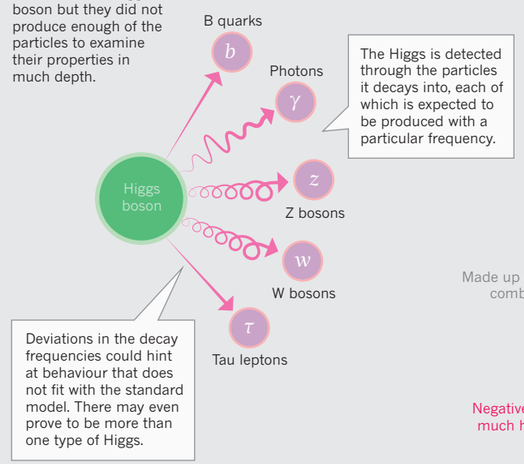
If the LHC makes supersymmetric particles, their lifetimes will be fleeting. But physicists can deduce their presence from the more-stable decay products. In at least one case, such SUSY clues could also be evidence for dark matter.



More collisions

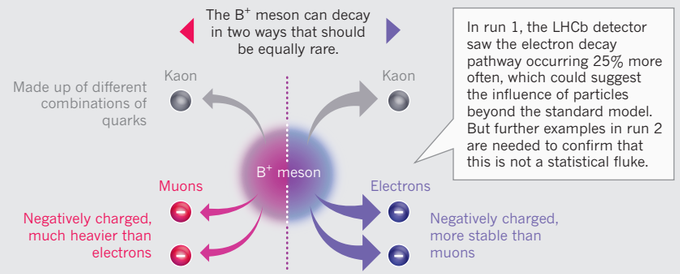
The Higgs factory

LHC experiments discovered the Higgs boson but they did not produce enough of the particles to examine their properties in much depth.



Known unknowns

More collisions will help to resolve some ongoing mysteries. One of these concerns an anomaly in the way a transient particle called a B^+ meson decays.



The human age

Momentum is building to establish a new geological epoch that recognizes humanity's impact on the planet. But there is fierce debate behind the scenes.

BY RICHARD MONASTERSKY

ILLUSTRATION BY JESSICA FORTNER

Almost all the dinosaurs have vanished from the National Museum of Natural History in Washington DC. The fossil hall is now mostly empty and painted in deep shadows as palaeobiologist Scott Wing wanders through the cavernous room.

Wing is part of a team carrying out a radical, US$45-million redesign of the exhibition space, which is part of the Smithsonian Institution. And when it opens again in 2019, the hall will do more than revisit Earth's distant past. Alongside the typical displays of *Tyrannosaurus rex* and *Triceratops*, there will be a new section that forces visitors to consider the species that is currently dominating the planet.

"We want to help people imagine their role in the world, which is maybe more important than many of them realize," says Wing.

This provocative exhibit will focus on the Anthropocene — the slice of Earth's history during which people have become a major geological force. Through mining activities alone, humans move more sediment than all the world's rivers combined. *Homo sapiens* has also warmed the planet, raised sea levels, eroded the ozone layer and acidified the oceans.

Given the magnitude of these changes, many researchers propose that the Anthropocene represents a new division of geological time. The concept has gained traction, especially in the past few years — and not just among geoscientists. The word has been invoked by archaeologists, historians and even gender-studies researchers; several museums

around the world have exhibited art inspired by the Anthropocene; and the media have heartily adopted the idea. "Welcome to the Anthropocene," *The Economist* announced in 2011.

The greeting was a tad premature. Although the term is trending, the Anthropocene is still an amorphous notion — an unofficial name that has yet to be accepted as part of the geological timescale. That may change soon. A committee of researchers is currently hashing out whether to codify the Anthropocene as a formal geological unit, and when to define its starting point.

But critics worry that important arguments against the proposal have been drowned out by popular enthusiasm, driven in part by environmentally minded researchers who want to highlight how destructive humans have become. Some supporters of the Anthropocene idea have even been likened to zealots. "There's a similarity to certain religious groups who are extremely keen on their religion — to the extent that they think everybody who doesn't practise their religion is some kind of barbarian," says one geologist who asked not to be named.

The debate has shone a spotlight on the typically unnoticed process by which geologists carve up Earth's 4.5 billion years of history. Normally, decisions about the geological timescale are made solely on the basis of stratigraphy — the evidence contained in layers of rock, ocean sediments, ice cores and other geological deposits. But the issue of the Anthropocene "is an order of magnitude more complicated than the stratigraphy", says Jan Zalasiewicz, a geologist at the University of Leicester, UK, and the chair of the Anthropocene Working Group that is evaluating the issue for the International Commission on Stratigraphy (ICS).

> "The geologic timescale, in my view, is one of the great achievements of humanity."

WRITTEN IN STONE

For geoscientists, the timescale of Earth's history rivals the periodic table in terms of scientific importance. It has taken centuries of painstaking stratigraphic work — matching up major rock units around the world and placing them in order of formation — to provide an organizing scaffold that supports all studies of the planet's past. "The geologic timescale, in my view, is one of the great achievements of humanity," says Michael Walker, a Quaternary scientist at the University of Wales Trinity St David in Lampeter, UK.

Walker's work sits at the top of the timescale. He led a group that helped to define the most recent unit of geological time, the Holocene epoch, which began about 11,700 years ago.

The decision to formalize the Holocene in 2008 was one of the most recent major actions by the ICS, which oversees the timescale. The commission has segmented Earth's history into a series of nested blocks, much like the years, months and days of a calendar. In geological time, the 66 million years since the death of the dinosaurs is known as the Cenozoic era. Within that, the Quaternary period occupies the past 2.58 million years — during which Earth has cycled in and out of a few dozen ice ages. The vast bulk of the Quaternary consists of the Pleistocene epoch, with the Holocene occupying the thin sliver of time since the end of the last ice age.

When Walker and his group defined the beginning of the Holocene, they had to pick a spot on the planet that had a signal to mark that boundary. Most geological units are identified by a specific change recorded in rocks — often the first appearance of a ubiquitous fossil. But the Holocene is so young, geologically speaking, that it permits an unusual level of precision. Walker and his colleagues selected a climatic change — the end of the last ice age's final cold snap — and identified a chemical signature of that warming at a depth of 1,492.45 metres in a core of ice drilled near the centre of Greenland[1]. A similar fingerprint of warming can be seen in lake and marine sediments around the world, allowing geologists to precisely identify the start of the Holocene elsewhere.

NATURE.COM
To hear more about the Anthropocene, visit:
go.nature.com/vybhfu

Even as the ICS was finalizing its decision on the start of the Holocene, discussion was already building about whether it was time to end that epoch and replace it with the Anthropocene. This idea has a long history. In the mid-nineteenth century, several geologists sought to recognize the growing power of humankind by referring to the present as the 'anthropozoic era', and others have since made similar proposals, sometimes with different names. The idea has gained traction only in the past few years, however, in part because of rapid changes in the environment, as well as the influence of Paul Crutzen, a chemist at the Max Plank Institute for Chemistry in Mainz, Germany.

Crutzen has first-hand experience of how human actions are altering the planet. In the 1970s and 1980s, he made major discoveries about the ozone layer and how pollution from humans could damage it — work that eventually earned him a share of a Nobel prize. In 2000, he and Eugene Stoermer of the University of Michigan in Ann Arbor argued that the global population has gained so much influence over planetary processes that the current geological epoch should be called the Anthropocene[2]. As an atmospheric chemist, Crutzen was not part of the community that adjudicates changes to the geological timescale. But the idea inspired many geologists, particularly Zalasiewicz and other members of the Geological Society of London. In 2008, they wrote a position paper urging their community to consider the idea[3].

Those authors had the power to make things happen. Zalasiewicz happened to be a member of the Quaternary subcommission of the ICS, the body that would be responsible for officially considering the suggestion. One of his co-authors, geologist Phil Gibbard of the University of Cambridge, UK, chaired the subcommission at the time.

Although sceptical of the idea, Gibbard says, "I could see it was important, something we should not be turning our backs on." The next year, he tasked Zalasiewicz with forming the Anthropocene Working Group to look into the matter.

A NEW BEGINNING

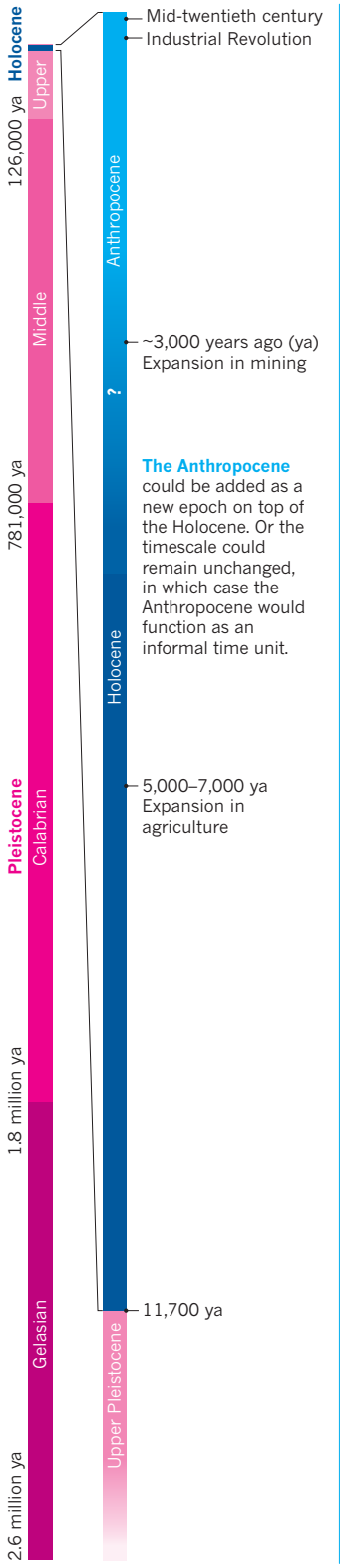
Since then, the working group has been busy. It has published two large reports ("They would each hurt you if they dropped on your toe," says Zalasiewicz) and dozens of other papers.

The group has several issues to tackle: whether it makes sense to establish the Anthropocene as a formal part of the geological timescale; when to start it; and what status it should have in the hierarchy of the geological time — if it is adopted.

When Crutzen proposed the term Anthropocene, he gave it the suffix appropriate for an epoch and argued for a starting date in the late eighteenth century, at the beginning of the Industrial Revolution. Between then and the start of the new millennium, he noted, humans had chewed a hole in the ozone layer over Antarctica, doubled the amount of methane in the atmosphere and driven up carbon dioxide concentrations by 30%, to a level not seen in 400,000 years.

When the Anthropocene Working Group started investigating, it compiled a much longer long list of the changes wrought by humans. Agriculture, construction and the damming of rivers is stripping away sediment at least ten times as fast as the natural forces of erosion. Along some coastlines, the flood of nutrients from fertilizers has created oxygen-poor 'dead zones', and the extra CO_2 from fossil-fuel burning has acidified the surface waters of the ocean by 0.1 pH units. The fingerprint of humans is clear in global temperatures, the rate of species extinctions and the loss of Arctic ice.

The group, which includes Crutzen, initially leaned towards his idea of choosing the Industrial Revolution as the beginning of the

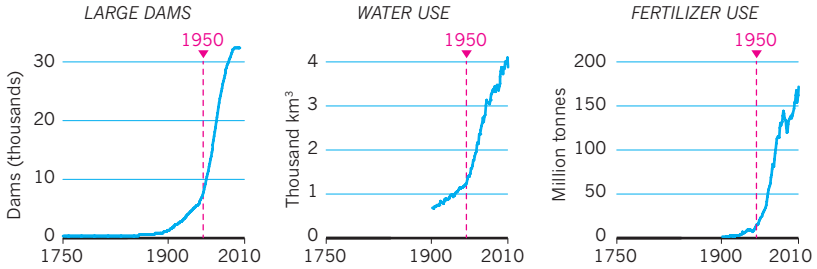


Humans at the helm

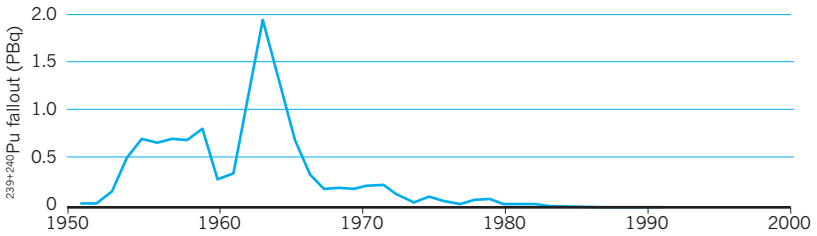
Researchers are studying whether the geological timescale should be modified to include the Anthropocene, a unit of time during which humans became a major force on the planet. Some support starting the Anthropocene in the mid-twentieth century, whereas others propose much earlier dates.

LATE-ANTHROPOCENE PROPOSAL

Human impacts on the environment surged in the mid-twentieth century, a trend visible in many records. That time has been called the Great Acceleration.

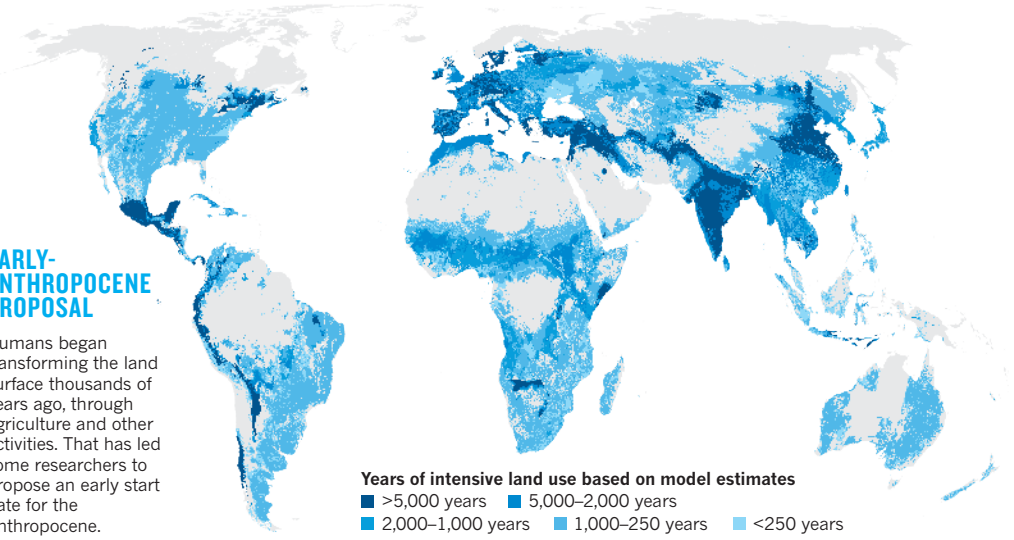


Radioactive fallout from nuclear blasts peaked in the mid-twentieth century, leaving a signal visible in sediments that has been proposed as a marker for the start of the Anthropocene.

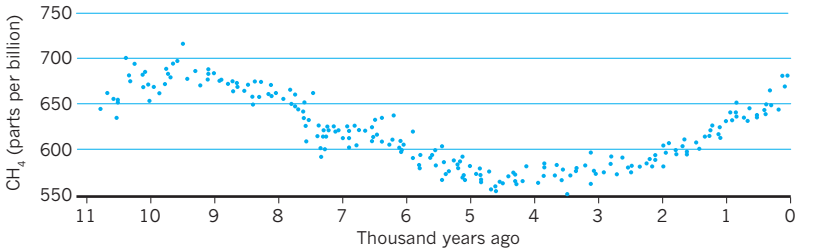


EARLY-ANTHROPOCENE PROPOSAL

Humans began transforming the land surface thousands of years ago, through agriculture and other activities. That has led some researchers to propose an early start date for the Anthropocene.



One potential stratigraphic marker is a rise in the atmospheric concentration of methane millennia ago, which is recorded in glacial ice. This could reflect increases in farming and animal herding.



SOURCES: DAMS/WATER/FERTILIZER, IGBP; FALLOUT, REF. 5; MAP, E. C. ELLIS *PHIL. TRANS. R. SOC. A* **369**, 1010–1035 (2011); METHANE, REF. 4

Anthropocene. But other options were on the table.

Some researchers have argued for a starting time that coincides with an expansion of agriculture and livestock cultivation more than 5,000 years ago[4], or a surge in mining more than 3,000 years ago (see 'Humans at the helm'). But neither the Industrial Revolution nor those earlier changes have left unambiguous geological signals of human activity that are synchronous around the globe.

This week in *Nature*, two researchers propose that a potential marker for the start of the Anthropocene could be a noticeable drop in atmospheric CO_2 concentrations between 1570 and 1620, which is recorded in ice cores (see page 171). They link this change to the deaths of some 50 million indigenous people in the Americas, triggered by the arrival of Europeans. In the aftermath, forests took over 65 million hectares of abandoned agricultural fields — a surge of regrowth that reduced global CO_2.

In the working group, Zalasiewicz and others have been talking increasingly about another option — using the geological marks left

by the atomic age. Between 1945 and 1963, when the Limited Nuclear Test Ban Treaty took effect, nations conducted some 500 above-ground nuclear blasts. Debris from those explosions circled the globe and created an identifiable layer of radioactive elements in sediments. At the same time, humans were making geological impressions in a number of other ways — all part of what has been called the Great Acceleration of the modern world. Plastics started flooding the environment, along with aluminium, artificial fertilizers, concrete and leaded petrol, all of which have left signals in the sedimentary record.

In January, the majority of the 37-person working group offered its first tentative conclusion. Zalasiewicz and 25 other members reported[5] that the geological markers available from the mid-twentieth century make this time "stratigraphically optimal" for picking the start of the Anthropocene, whether or not it is formally defined. Zalasiewicz calls it "a candidate for the least-worst boundary".

The group even proposed a precise date: 16 July 1945, the day of the first atomic-bomb blast. Geologists thousands of years in the future would be able to identify the boundary by looking in the sediments for the signature of long-lived plutonium from mid-century bomb blasts or many of the other global markers from that time.

A MANY-LAYERED DEBATE

The push to formalize the Anthropocene upsets some stratigraphers. In 2012, a commentary published by the Geological Society of America[6] asked: "Is the Anthropocene an issue of stratigraphy or pop culture?" Some complain that the working group has generated a stream of publicity in support of the concept. "I'm frustrated because any time they do anything, there are newspaper articles," says Stan Finney, a stratigraphic palaeontologist at California State University in Long Beach and the chair of the ICS, which would eventually vote on any proposal put forward by the working group. "What you see here is, it's become a political statement. That's what so many people want."

> "It's become a political statement. That's what so many people want."

Finney laid out some of his concerns in a paper[7] published in 2013. One major question is whether there really are significant records of the Anthropocene in global stratigraphy. In the deep sea, he notes, the layer of sediments representing the past 70 years would be thinner than 1 millimetre. An even larger issue, he says, is whether it is appropriate to name something that exists mainly in the present and the future as part of the geological timescale.

Some researchers argue that it is too soon to make a decision — it will take centuries or longer to know what lasting impact humans are having on the planet. One member of the working group, Erle Ellis, a geographer at the University of Maryland, Baltimore County, says that he raised the idea of holding off with fellow members of the group. "We should set a time, perhaps 1,000 years from now, in which we would officially investigate this," he says. "Making a decision before that would be premature."

That does not seem likely, given that the working group plans to present initial recommendations by 2016.

Some members with different views from the majority have dropped out of the discussion. Walker and others contend that human activities have already been recognized in the geological timescale: the only difference between the current warm period, the Holocene, and all the interglacial times during the Pleistocene is the presence of human societies in the modern one. "You've played the human card in defining the Holocene. It's very difficult to play the human card again," he says.

Walker resigned from the group a year ago, when it became clear that he had little to add. He has nothing but respect for its members, he says, but he has heard concern that the Anthropocene movement is picking up speed. "There's a sense in some quarters that this is something of a juggernaut," he says. "Within the geologic community, particularly within the stratigraphic community, there is a sense of disquiet."

Zalasiewicz takes pains to make it clear that the working group has not yet reached any firm conclusions. "We need to discuss the utility of the Anthropocene. If one is to formalize it, who would that help, and to whom it might be a nuisance?" he says. "There is lots of work still to do."

Any proposal that the group did make would still need to pass a series of hurdles. First, it would need to receive a supermajority — 60% support — in a vote by members of the Quaternary subcommission. Then it would need to reach the same margin in a second vote by the leadership of the full ICS, which includes chairs from groups that study the major time blocks. Finally, the executive committee of the International Union of Geological Sciences must approve the request.

At each step, proposals are often sent back for revision, and they sometimes die altogether. It is an inherently conservative process, says Martin Head, a marine stratigrapher at Brock University in St Catharines, Canada, and the current head of the Quaternary subcommission. "You are messing around with a timescale that is used by millions of people around the world. So if you're making changes, they have to be made on the basis of something for which there is overwhelming support."

Some voting members of the Quaternary subcommission have told *Nature* that they have not been persuaded by the arguments raised so far in favour of the Anthropocene. Gibbard, a friend of Zalasiewicz's, says that defining this new epoch will not help most Quaternary geologists, especially those working in the Holocene, because they tend not to study material from the past few decades or centuries. But, he adds: "I don't want to be the person who ruins the party, because a lot of useful stuff is coming out as a consequence of people thinking about this in a systematic way."

If a proposal does not pass, researchers could continue to use the name Anthropocene on an informal basis, in much the same way as archaeological terms such as the Neolithic era and the Bronze Age are used today. Regardless of the outcome, the Anthropocene has already taken on a life of its own. Three Anthropocene journals have started up in the past two years, and the number of papers on the topic is rising sharply, with more than 200 published in 2014.

By 2019, when the new fossil hall opens at the Smithsonian's natural history museum, it will probably be clear whether the Anthropocene exhibition depicts an official time unit or not. Wing, a member of the working group, says that he does not want the stratigraphic debate to overshadow the bigger issues. "There is certainly a broader point about human effects on Earth systems, which is way more important and also more scientifically interesting."

As he walks through the closed palaeontology hall, he points out how much work has yet to be done to refashion the exhibits and modernize the museum, which opened more than a century ago. A hundred years is a heartbeat to a geologist. But in that span, the human population has more than tripled. Wing wants museum visitors to think, however briefly, about the planetary power that people now wield, and how that fits into the context of Earth's history. "If you look back from 10 million years in the future," he says, "you'll be able to see what we were doing today." ■ SEE EDITORIAL P.129

Richard Monastersky *is a features editor for* Nature *in Washington DC.*

1. Walker, M. *et al. J. Quat. Sci.* **24,** 3–17 (2009).
2. Crutzen, P. J. & Stoermer, E. F. *IGBP Newsletter* **41,** 17–18 (2000).
3. Zalasiewicz. J. *et al. GSA Today* **18**(2), 4–8 (2008).
4. Ruddiman, W. F. *Ann. Rev. Earth. Planet. Sci.* **41,** 45–68 (2013).
5. Zalasiewicz, J. *et al. Quatern. Int.* http://dx.doi.org/10.1016/j.quaint.2014.11.045 (2015).
6. Autin, W. J. & Holbrook, J. M. *GSA Today* **22**(7), 60–61 (2012).
7. Finney, S. C. *Geol. Soc. Spec. Publ.* **395,** 23–28 (2013).

WARS WITHOUT END

The world is full of bloody conflicts that can drag on for decades. Some researchers are trying to find resolutions through complexity science.

BY DAN JONES

In the seven decades that Colombia has been riven by civil war, the country has seen kidnappings, rapes, terrorist attacks and pitched battles that have cost more than 220,000 lives and displaced millions of people. Negotiations, peace accords and ceasefires have come and gone to little lasting effect.

The latest round of this seemingly unending cycle began in August 2012, when the Marxist rebels of the Revolutionary Armed Forces of Colombia (FARC) agreed to meet with the central government in yet another round of peace talks. But the negotiations collapsed in November after the rebels kidnapped a Colombian army general. The talks have since resumed, but even if they one day yield a peace accord, there is no guarantee it will hold. More than one-third of the world's peace agreements and ceasefires since the 1950s have relapsed into violence within five years.

Colombia's long history of strife is a classic example of 'intractable' conflict — a self-perpetuating cycle of hostility that can grind on for decades. Such conflicts are relatively scarce — only about 5% of the world's myriad wars qualify — but their longevity means that they exert a huge toll on societies. Their tragic poster child is the 68-year-long Israeli–Palestinian conflict. But the list also includes India and Pakistan's equally long battle over Kashmir, and Sri Lanka's 26-year civil war. The Democratic Republic of the Congo (DRC) has been riven by conflict since 1996, as has South Sudan since its inception in 2011. Any number of intractable conflicts may now be emerging in the Middle East as Libya, Syria and Iraq are ripped apart by sectarian violence and with the rise of the Islamist group ISIS (see 'Intractable conflicts'). The intensifying civil war in eastern Ukraine

The Israeli–Palestinian conflict has been ongoing for 68 years.

ASHRAF AMRA/ANADOLU AGENCY/GETTY IMAGES

may eventually join the list as well.

By definition, these are the conflicts that are resistant to all the mainstream techniques of dispute resolution, says Robert Ricigliano, a mediation expert at the University of Wisconsin Milwaukee. Typically they are plagued by a history of "fixes that fail", he says — peace agreements that collapse within days or weeks. "We mediate agreements, change leaders, arbitrate boundaries," he says. "But those things don't necessarily get at the underlying dynamics fuelling conflict."

He and a growing chorus of other conflict researchers have therefore been pushing for a fresh approach — one that views intractable conflicts as dynamic, complex systems similar to cells, ant colonies or cities, and analyses them with the mathematical and computational tools developed over the past 30 years in complexity science.

Mainstream practitioners tend to be dubious, says Dan Smith, head of the London-based peace-building organization International Alert. "We know that conflicts are complex," he says. "What would be useful would be a clearer idea of what to do about it."

But Ricigliano and others have begun to answer that criticism by using complexity-inspired techniques to help resolve conflicts in places such as the DRC. They say that the approach can be a much-needed corrective to business as usual in the conflict-resolution world, where governments and international organizations too often tackle conflicts piecemeal. These bodies tend to "look at the economy, or governance, or gender relations or education as if each existed in isolation", says Smith. "It's a convenient way to handle the issues, but it means you don't really address the complex reality."

HARD PROBLEMS

It was just this kind of blinkered thinking that led psychologist Peter Coleman to rebel. It was 2000, recalls Coleman, head of the Morton Deutsch International Center for Cooperation and Conflict Resolution at Columbia University in New York City. He had broken his foot and decided to spend his convalescence at home delving into the research literature on intractable conflict. But what he found left him deeply frustrated. "People had their simple, sovereign theories about why conflicts become intractable," he says. "It's because of trauma, or social identity or a history of humiliation. We understood pieces of the problem, but not how they interact."

Coleman discovered an alternative approach just a few years later, when he came across the work of social psychologists Robin Vallacher and Andrzej Nowak, both now at Florida Atlantic University in Boca Raton. Their work was not directly related to conflict — they were studying things such as how the human sense of self emerges, and how feelings about others can switch from positive to negative. But Coleman was impressed with Vallacher and Nowak's use of a mathematical tool known as dynamical systems theory to analyse their results.

> "SUCCESS DOESN'T MEAN THAT WE'VE ENDED THE CONFLICT. IT MEANS WE'VE ENGAGED A SYSTEM SO THAT VIOLENCE DECLINES OVER TIME."

Made famous by James Gleick's 1987 book *Chaos*, this theory provides a framework for understanding a remarkably broad range of complex systems, from weather patterns to neural activity in the brain. One way to visualize the mathematics is to imagine a landscape of hills and valleys. The behaviour of the complex system corresponds to the path of a ball rolling across this landscape. The trajectory becomes very complicated as the ball is deflected by the hills. But eventually, the ball will get trapped in one of the valleys, where it will either cycle endlessly around the walls or sink to the middle and lie still. The ball's final trajectory or resting place is called an attractor.

To Coleman, this kind of entrapment was the perfect metaphor for the stable, if destructive, patterns of social behaviour seen in intractable conflicts. The landscapes in this case are mainly psychological and social, comprising innumerable strata of history, identity and collective memories of harms suffered at the hands of the 'other'. Yet the resulting conflict attractors are terribly real, he says, with psychological forces conspiring to "create simplistic narratives about conflicts that are devoid of nuance and keep us locked in".

To make this mathematical view of intractable conflicts into something more than a metaphor — and hopefully to turn it into a set of tools that could make a difference in the real world — Coleman, Vallacher and Nowak in 2004 formed the Dynamics of Conflict working group, which has since attracted four more members.

As a result of this collaboration, Nowak has started to create computational models that capture the dynamics of conflicts. These include 'agent-based' simulations that contain thousands of digital robots — the agents — each of which embodies some of the simple behaviours that social psychologists believe have a role in conflict. One such model, developed with researchers outside the working group, features agents that vary in how competitive or cooperative they are, and adjust those proclivities according to how much hostility or aggression they experience from the other agents.

In this simulation[1], small conflagrations flare up and die down much as they do in real communities. Occasionally, however, the conflicts expand until they lock the whole virtual community into a cycle of recrimination — the classic sign of intractability. Working with Dean Pruitt of the School for Conflict Analysis and Resolution at George Mason University in Arlington, Virginia, Nowak has also developed mathematical models showing how attractors can explain the escalation of conflicts that tips them into an intractable state[2]. Now he and his colleagues are working on the next step: comparing the evolution of communities in these simple models with data from real-world conflicts such as the Israeli–Palestinian stand-off. "This is the first time we've added empirical data to a dynamical model, and we're getting promising results," says Nowak.

MAKING SENSE OF THE SYSTEM

Another line of research is to move from generalities to specifics, and develop visualization tools that can help mediators to untangle the complexities of real-world conflicts. The hope is that such 'conflict maps' will help researchers to keep track of the interconnections between players and events, and make clear the feedback loops and key networks that can escalate or inhibit conflict.

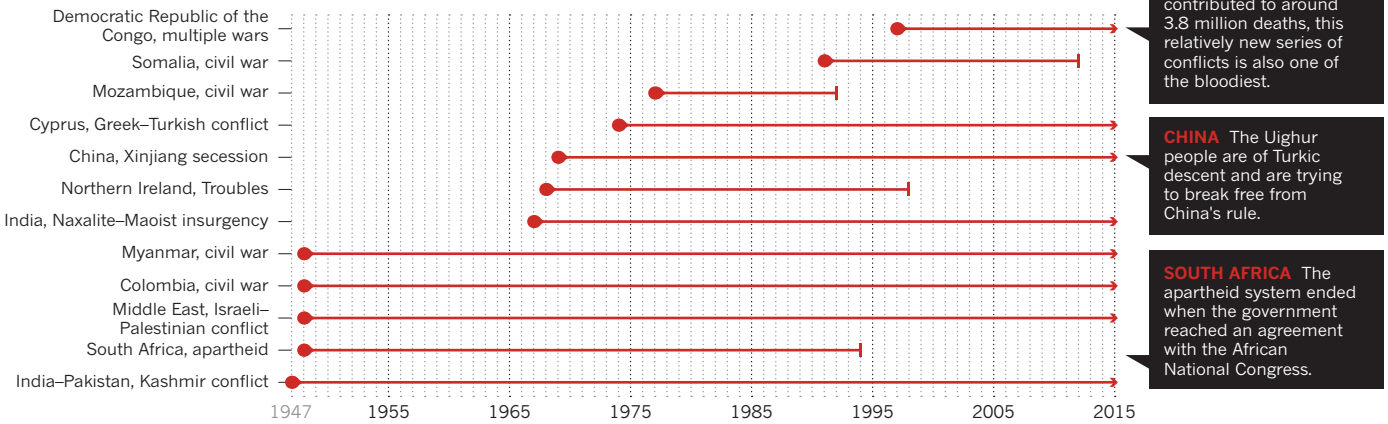
Conflict maps can take many forms, from hand-drawn sketches on a whiteboard to computer-generated networks based on real data. But whatever their form, they get strong endorsement from Ricigliano, who has worked on peace-building interventions in areas ranging from Colombia and South Africa to Iraq and Cambodia.

In 2000, for example, Ricigliano went to the DRC to try to find some resolution to the Second Congo War: a blood-drenched conflict between various rebel groups and Mai-Mai militias fighting for the government. Behind the scenes, he and his colleagues watched the unravelling of one hard-won peace agreement after another. "At best we were having a neutral impact," he says, "and maybe even a negative one."

But then in 2002, he and his colleagues began to map all the connections between warring parties and competing interests in the conflict. The maps made it clear that local groups were being manipulated by national rebel organizations, who wanted the conflict to continue because it allowed them to access valuable minerals. "So we shifted tactics, and began trying to break the links between the national-level actors who were manipulating local actors, and to facilitate local-level ceasefires of significance, says Ricigliano.

INTRACTABLE CONFLICTS

These maelstroms of grievance and mistrust can go on and on. Below is a sample of those that have emerged since the Second World War.



By 2003, these dialogues had helped the United Nations to negotiate a transition government that included the major rebel groups, and violence declined. "It wasn't perfect," says Steve Smith, an independent conflict-resolution consultant who was in the DRC at the time. "Not everyone was in agreement, and little conflicts continued, but we had a structure in place and a direction to go."

STATE OF MIND

Beyond the models and the maps, advocates of systems thinking are hoping to spread a shift in perspective on intractable conflicts. One convert is Andrea Bartoli, dean of the School of Diplomacy and International Relations at Seton Hall University in South Orange, New Jersey, and a mediator who has worked in countries such as Mozambique and Kosovo. When he first learned about the dynamical systems perspective in discussions with Coleman a little over a decade ago, he says, "it provided a new language for talking about conflict, and opened up new ways to think about old problems". He has since joined the Dynamics of Conflict working group, and in 2009 joined with Coleman and Beth Yoshida-Fisher, director of the Negotiation and Conflict Resolution program at Columbia University, to set up the Advanced Consortium on Cooperation, Conflict, and Complexity (AC4) there.

That new language can be a revelation even to professionals, says Naira Musallam, a conflict researcher at New York University's Center for Global Affairs and a member of both the Dynamics of Conflict group and AC4. She tells the story of a course she teaches at the US Military Academy West Point in New York, in which she starts by running through a list of common mental shortcomings in how people think about conflict, poverty and other social problems. "We compare fluid situations to fixed things," she says, "we think in straight lines rather than loops, we focus on understanding problems and assume that this will lead to solutions, and often miss the unintended consequences of well-intentioned interventions."

After one class, says Musallam, an officer who had served in Iraq and Afghanistan wrote to her. "I know many good people who have died because of errors [highlighted] on this list," he wrote. "I also see several errors that I have made before … It's frustrating that this is the first time that I've seen this list in a way that challenges my world view around conflict."

These same straight-line assumptions are also built into the way in which many institutions operate, says Musallam — and not just those devoted to peace-building. "They want nice, tidy plans for interventions, and clear deliverables over the short term", she says. This often leads to plans to 'solve' complex problems through a series of discrete steps that are defined in advance by experts.

One of the key lessons of the systems mindset is to stop approaching conflicts as problems that need to be fixed, says Ricigliano, and instead think of them as systems with underlying dynamics that need to shift. "Success doesn't mean that we've ended the conflict," he says. "It means we've engaged a system so that violence declines over time."

This view is finding increasing support from outside allies. The non-profit Berlin-based Berghof Foundation, for example, has used systems thinking in its efforts to resolve political and ethnic violence in countries such as Sri Lanka, which has been torn by civil war since 1983.

But there is plenty of room for scepticism. Dan Smith, for one, is sympathetic to the complex systems view of conflict, but is wary of its sweeping generalizations. "Any analysis employing these principles is only going to be as good as the analyst doing it," he says. "You can have the best methodology, but if you have an uninformed or incurious analyst, you won't get good results."

Even advocates admit that specific recommendations are a work in progress. That is why in 2013, Coleman and Ricigliano joined with others to set up an annual five-day workshop known as the Dynamic Systems Theory Innovation Lab, which brings together biologists, economists, physicists, political scientists and other scholars and practitioners to talk about real-world applications. "We hope that five years out, we'll have a better idea of what matters most," says Coleman.

There is already a growing body of experiments they can draw on. In Israel, for example, a series of anti-conflict interventions being developed under the leadership of psychologist Eran Halperin at the Interdisciplinary Center Herzliya in Israel have proved effective in making people more open to seeing things from the other side's point of view[3–5].

Although the label 'intractable conflict' implies unending strife, no struggle lasts forever. As the 1980s drew to a close, South Africa had been locked in racial conflict for decades and was on the brink of a civil war between increasingly militant members of the African National Congress (ANC) and the government of President Frederik Willem De Klerk. Amid international condemnation of the apartheid system, and fearing that the country could become engulfed in a bloody street war, De Klerk began releasing imprisoned ANC members in late 1989. Finally, in February 1990, he freed ANC leader Nelson Mandela after 27 years in prison. That conciliatory move was the tipping point for the emergence of multiracial democracy within three years.

South Africa's long transition was a difficult journey, with many losses and setbacks along the way — par for the course for any intractable conflict. Yet as Mandela once famously said: "It always seems impossible until it's done." ■

Dan Jones *is a freelance writer in Brighton, UK.*

1. Nowak, A., Deutsch, M., Bartkowski, W. & Solomon, S. *Peace Confl.* **16,** 189–209 (2010).
2. Pruitt, D. G. & Nowak, A. *Int. J. Confl. Manage.* **25,** 387–406 (2014).
3. Hameiri, B., Porat, R., Bar-Tal, D., Bieler, A. & Halperin, E. *Proc. Natl Acad. Sci. USA* **111,** 10996–11001 (2014).
4. Halperin, E., Russell, A. G., Trzesniewski, K. H., Gross, J. J. & Dweck, C. S. *Science* **333,** 1767–1769 (2011).
5. Nasie, M., Bar-Tal, D., Pliskin, R., Nahhas, E. & Halperin, E. *Pers. Soc. Psychol. Bull.* **40,** 1543–1556 (2014).

COMMENT

PORNCHAI KITTIWONGSAKUL/AFP/GETTY

A Tuareg woman carries water through a sandstorm in drought-ridden Mali.

Put people at the centre of global risk management

An individual focus is needed to assess interconnected threats and build resilience worldwide, urge **Jan Willem Erisman** and colleagues.

Globalization is changing the nature of risk. Natural and social systems — from climate to energy, food, water and economies — are tightly coupled. Abrupt changes in one have a domino effect on others. Floods in Thailand in 2010, for example, led to a global shortage of computer hard disks as a result of factories closing, as well as more than US$330 million in damage and around 250 deaths.

The exposure of people and assets to risks is increasing worldwide. From 1980 to 2012, annual economic losses from environmental disasters rose more than sevenfold, from about $20 billion to $150 billion a year[1].

Yet most risk assessments ignore networked threats[2,3]. The annual Global Risks report of the World Economic Forum considers risks qualitatively, based on the views of experts[4]. But global outlooks remain sectorial and too coarse to guide individuals, organizations, municipalities or nations.

Risk reports also neglect the collective impacts of personal choices[3]. For example, eating more beef causes deforestation and biodiversity loss in the Amazon. Local dams for hydropower or water storage alter sediment flows to fertile coastal regions. The movement of people from the ▶

Global risks and losses from extreme weather, as in Thailand's 2010 floods, and agricultural failures, such as locusts destroying crops in Madagascar, are rising.

countryside to cities affects water, food, climatic and energy systems planet-wide.

Understanding networked risks is essential for achieving the United Nations Sustainable Development Goals, which are being defined this year[5]. The 17 proposed goals are interdependent. For example, the stimulation of renewable energies and biofuels to address climate change also affects food production and water resources.

BROAD FOCUS

We propose a systems-based approach for quantifying risk that integrates individual responses and considers the transfer of information and feedback mechanisms across networks (see 'Safety secured'). Such an approach identifies pinch points — geographic, economic and social — so that key systems and individual behaviours can be made more sustainable and resilient.

Current global-change risk assessments take a top-down approach and target single stressors, such as the climate. They focus on the most vulnerable and at-risk communities, infrastructure, sectors, ecosystems and areas. Links between extreme weather and climate change have begun to be addressed, but wider impacts on land degradation, food and energy production, water supply and environmental hazards have not.

Disaster-reduction frameworks, such as the UN Post-2015 Development Agenda, which will be agreed this month in Japan, aim to improve reactions to adverse events once they have happened. But the UN agenda does not promote resilience in general or help stakeholders such as farmers or municipal leaders to manage multiple risks.

Programmes for delivering knowledge about risk to sectors of society are too narrow. Climate services inform the agriculture and insurance sectors about climate change. But their academic focus does not serve corporate clients, who want climate data packaged into products that they can use to manage, for example, exposure to market disruptions or rising energy prices.

The Climate Corporation in San Francisco, California, sells weather and agronomic data-monitoring and modelling tools to farmers. But its products do not, for instance, consider other impacts such as the risks of air and water pollution associated with the use of nitrogen fertilizers[6].

Approaches to communicating a broader set of global risks appeal to researchers and policy-makers. For example, the 'planetary boundaries' concept[7] identifies tipping points in nine key Earth systems (including climate change, biodiversity and the nitrogen cycle) above which Earth's habitability would be threatened. But global limits are difficult to translate into targets or strategies that are meaningful for a particular company, city or region.

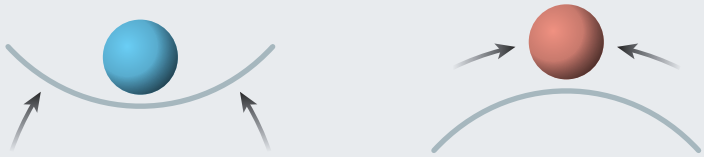
How then should scientists, insurance companies, policy-makers and other stakeholders combine risk assessments across scales, stressors and sectors?

USER FIRST

We argue that Earth-system risk management should follow the example of health-care systems, in which emphasis is switching from medicalization to supporting people's ability to adapt and self-manage[8]. Collectively, individual choices feed back into the community and help it to lower its health risks.

SAFETY SECURED

Promoting overall resilience (left) rather than managing many individual risks (right) is the best way to minimize impacts from adverse events.



RESILIENCE
- Concerns whole system
- Aims for long-term security
- Requires indirect management
- Self-regulating
- Makes use of variability
- Seeks dynamic equilibrium

RISK MANAGEMENT
- Focuses on single risks
- Aims for short-term security
- Requires direct intervention
- Needs continuous monitoring
- Eliminates variability
- Seeks static equilibrium

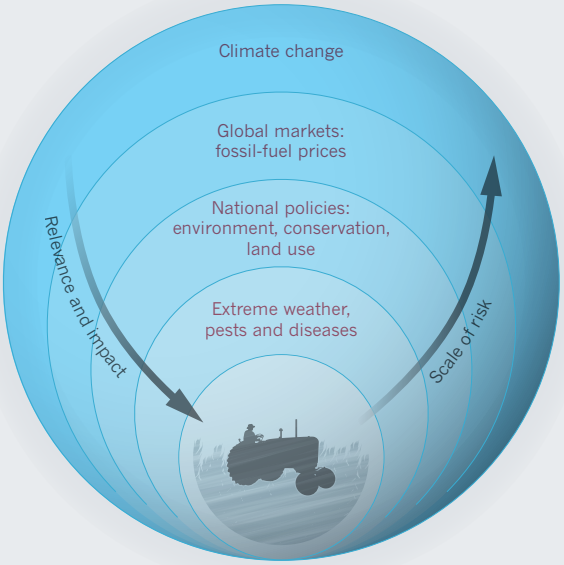
SOURCE: J. TEN NAPEL, F. BIANCHI & M. W. P. BESTMAN IN *INVENTION FOR A SUSTAINABLE DEVELOPMENT OF AGRICULTURE* 32–53 (2006)

LEFT: PORNCHAI KITTIWONGSAKUL/AFP/GETTY; RIGHT: BILAL TARABEY/AFP/GETTY

SOURCE: J.W.E.

NETWORKED THREATS

As well as immediate risks such as droughts and floods, individuals should factor in remote threats such as climate change into their decisions. If risks from the local to the global and connections between them are assessed, people can choose effective actions that build resilience.



Risk management must therefore start with the users — be they people, organizations, municipalities or nations. Risks should be identified and prioritized in expanding circles around the user (see 'Networked threats'), from local and short-term risks to more distant and long-term related global threats.

Take food. Supplies are threatened by elevated production costs, ecosystem, water and soil-quality impairment, food wastage and nutrient losses, poor food distribution and alienation of consumers from producers. Yet farmers consider only immediate factors — maximizing yields, avoiding disease and short-term price fluctuations — when deciding how and when to plant crops.

In our approach, farmers would also consider climate change, energy prices, floods and droughts and ecosystem services. The wider ecological and social repercussions of personal decisions such as whether to use more fertilizer or pesticides, expand soil tillage or irrigation would become more apparent. Worldwide, 10% of farmers manage 70% of the agricultural land, so the side effects of such localized choices can be widespread.

RADICAL REFRAMING

In practical terms, a networked risk-assessment model should combine standard techniques for individual risk assessments (such as those set out for enterprises by the International Organization for Standardization) with a mechanism to capture the complexities of human behaviour. One such method is agent-based modelling[9], which uses simulations of a collection of computational entities that interact according to a set of mathematical rules. This approach has been used to model stock-market trends, traffic flows and the spread of epidemics.

Two major shifts in thinking are needed to deliver the global risk-network model. First, the risk narrative needs to be reframed to put the individual at the centre. Second, risk modelling should adapt to take a broad focus — encompassing environmental and socio-economic risks across the whole Earth system.

The UN's sustainability and disaster-reduction programmes should adopt this user-centric focus and redirect their existing efforts. The UN-led Global Framework for Climate Services should be similarly extended to include inventories of issues that matter to the individual (collated through platforms such as the UN website vote.myworld2015.org).

Relevant risks at particular scales will need to be defined and methods for analysing them jointly developed. Future Earth, a global research hub launching this year to provide the knowledge and support to accelerate transformations to a sustainable world, should coordinate the research.

Partnerships must be built across disciplines to supply and share data and analysis tools. Practitioners from the private and public sectors will need to work with economists, engineers, social scientists, information specialists and climate and Earth-system experts.

Investment by public–private partnerships will be essential to amass the necessary resources, maximize uptake of this multiscale approach, stimulate innovation from industry and guarantee that the user's needs are at the core. As the cost of disasters increases each year, the impetus for both governments and industry to invest in risk management and resilience is clear. ■

Jan Willem Erisman *is director of the Louis Bolk Institute in Driebergen, the Netherlands.* **Guy Brasseur** *is a senior scientist at the Max Planck Institute for Meteorology in Hamburg, Germany.* **Philippe Ciais** *is a senior researcher at the Laboratory for Climate Sciences and the Environment, University of Versailles, France.* **Nick van Eekeren** *is senior researcher at the Louis Bolk Institute in Driebergen, the Netherlands.* **Thomas L. Theis** *is director of the Institute for Environmental Science and Policy, University of Illinois at Chicago, USA.*
e-mail: j.erisman@louisbolk.nl

1. Guha-Sapir, D., Hoyois, P. & Below, R. *Annual Disaster Statistical Review 2013: The Numbers and Trends* (Centre for Research on the Epidemiology of Disasters, 2014).
2. Helbing, D. *Nature* **497,** 51–59 (2013).
3. Adger, W. N., Eakin, H. & Winkels, A. *Front. Ecol. Environ.* **7,** 150–157 (2009).
4. World Economic Forum. *Global Risks 2015* (WEF, 2015).
5. Open Working Group of the General Assembly on Sustainable Development Goals. *Open Working Group Proposal for Sustainable Development Goals* (United Nations, 2014).
6. Erisman, J. W., Sutton, M. A., Galloway, J., Klimont, Z. & Winiwarter, W. *Nature Geosci.* **1,** 636–639 (2008).
7. Rockström, J. *et al. Nature* **461,** 472–475 (2009).
8. Huber, M., Knottnerus, J. A., Green L. *et al. Br. Med. J.* **343,** 235–237 (2011).
9. Gilbert, N. *Agent-based Models* (Sage, 2007).

OWEN FRANKEN/CORBIS

Learning English is essential for modern scientists — but German and French were once more significant.

LINGUISTICS

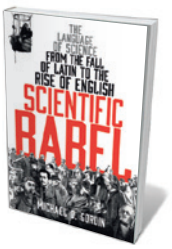
The ascent of English

Andrew Robinson salutes a chronicle of how one language came to dominate science.

A scientific paper published in 1905 gloried in the title *Zur Elektrodynamik bewegter Körper*. Today, Albert Einstein's 'On the electrodynamics of moving bodies', which introduced the special theory of relativity, would be published in English. English has become the language of almost every leading journal across the natural sciences, whatever its country of origin. Large conferences held in non-anglophone countries, such as those of the European Geosciences Union, often use English. Of the major producers of scientific research, only China and, to a lesser extent, Japan host international conferences in their own languages.

In 1905, however, some 30% of global scientific literature was in German, with a similar proportion in English, marginally less in French and much less in Russian and Japanese. So reveals US historian Michael Gordin in *Scientific Babel*, a massive, erudite and engaging study of the role of languages in science based on 15 years of research — and drawing on Gordin's knowledge of French, German, Russian, Esperanto and Latin. The numerous translations are generally his own.

The dominance of English — unpredicted a century ago — is rooted in Germany's defeat in the First World War. For some years afterwards, there was an international boycott of German scientists and attempts were made to curb the use of German by the League of Nations and 22 US states. The advent of the Third Reich in 1933 boosted English as the scientific lingua franca, as did the United States' postwar ascendancy in scientific output and geopolitical power — along with a perception of English as neutral.



Scientific Babel: The Language of Science from the Fall of Latin to the Rise of English
MICHAEL GORDIN
Profile/Univ. Chicago Press: 2015.

Gordin asks, with a touch of irony, whether this English-language "fait accompli" is always good for science. Although he finds that most scientists are in principle inclined to embrace the idea of one language for communicating, the dominance of English can disadvantage non-English speakers. The most creative thinking tends to be done in the language in which a person feels most at home. As Fields Medal winner Laurent Lafforgue noted (in French) in 2005: "it is to the degree that the French mathematical school remains attached to French that it conserves its originality and its force".

Gordin asks: does history suggest a future alternative? He considers relevant historical episodes in detail. Latin, for example, became the language of European science during the Italian Renaissance, but its use began to decline in the seventeenth century. Thus, Galileo Galilei turned to Italian, and Isaac Newton shifted from Latin for his *Principia Mathematica* (1687) to English for his *Opticks* (1704). During the Enlightenment, European libraries collected roughly one-third of their books in Latin, one-third in French and the rest in the local vernacular. Barring taxonomic nomenclature, the use of Latin had died out among leading scientists by the time of Charles Darwin, who wrote in English.

The linguistic complexity in science in the late nineteenth century is demonstrated by the story of the periodic table and its contested origin, which Gordin explored in his 2004 book *A Well-Ordered Thing* (Basic Books). When the German-language journal *Zeitschrift für Chemie* mistranslated an 1869 Russian abstract by Dmitri Mendeleev, a vehement priority dispute blew up between Mendeleev and German chemist Lothar Meyer. In a crucial sentence, "The elements ordered according to the magnitude of their atomic weights show a periodic change in properties", a rushed translator used the

German word *stufenweise* ('phased') instead of *periodische* ('periodic'); as a result, Meyer claimed precedence for his own research. When Mendeleev objected, Meyer replied: "It seems to me an excessive demand that we German chemists read, besides those articles appearing in the German and Romance languages, also those in the Slavic languages". He did not mention English.

By the end of the nineteenth century, scientists everywhere were obsessed with a multilingual information overload — Gordin's scientific babel. The solution seemed to be an auxiliary universal language. Volapük ('Worldspeak') was invented in 1880; the better-known Esperanto arose in 1887, and its offshoot, Ido, arrived in 1907. Gordin sympathetically analyses these artificial languages — taken seriously by leading scientists of the time — through the lens of Ido advocate Wilhelm Ostwald, a Nobel-prizewinning German chemist. In-fighting dissolved the movement, and Ostwald abandoned Ido during the First World War, championing German as an international language.

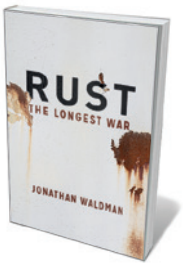
During the cold war, and especially after the Soviet Union launched Sputnik in 1957, much scientific attention switched to literature in Russian, which by 1970 reached 20% of the global output. In 1961, 85 Soviet journals were being translated into English, with US government funding. Preposterous claims were made for machine translation from Russian into English. Both translation programmes were eventually abandoned in favour of increased Russian-language teaching for US scientists — until the 1991 collapse of the Soviet Union sealed the fate of scientific Russian beyond its own borders. A lively Russian-language journals scene still prevails in Russia.

"By the end of the nineteenth century, scientists everywhere were obsessed with a multilingual overload."

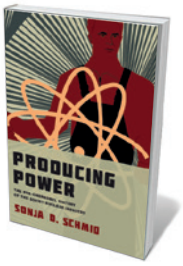
Anglophone dominance is unlikely to change soon, says Gordin. If scientific importance were based on population, Spanish would be a major scientific language; if on geopolitical power, scientists would publish much more in Chinese. In the 1660s and later, philosopher and mathematician Gottfried Leibniz advocated a universal writing system for science independent of any spoken language, similar to mathematical notation. This must stay a dream: intellectual activity demands language. As the polyglot Gordin concludes, "we remain bound to the constraints of history, to the shackles of the words in human languages: untranslatable yet intelligible, frustrating yet infinitely beguiling". ■

Andrew Robinson *is the author of* The Story of Writing.
e-mail: andrew.robinson33@virgin.net

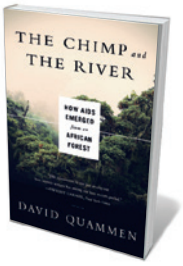
Books in brief



Rust: The Longest War
Jonathan Waldman SIMON AND SCHUSTER *(2015)*
Corrosion has killed people in nuclear power plants, taken out planes in mid-air and reddened the face of Mars. So notes environmental journalist Jonathan Waldman in this dexterous technological study of this insidious process, which is nibbling away at Western civilization. The science compels, but what leap from the page are Waldman's snapshots of rust geeks — such as the team that rebuilt the hole-ridden metal skin of New York's Statue of Liberty in the 1980s, and Bhaskar Neogi, 'integrity manager' of the Trans-Alaska Pipeline System, one of the heftiest metal objects in the Western Hemisphere.

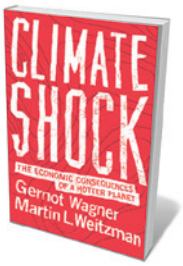


Producing Power: The Pre-Chernobyl History of the Soviet Nuclear Industry
Sonja D. Schmid MIT PRESS *(2015)*
In the annals of nuclear meltdown, the April 1986 explosion at Chernobyl in Soviet Ukraine remains the most devastating, contaminating thousands of square kilometres of land. This trenchant study by science historian Sonja Schmid digs deep into the catastrophe's tangled prehistory to make nuanced sense of it. She unravels key scientific, social and political factors, from the plant's lack of 'redundant' safety features to rivalries in the Soviet nuclear industry and inefficiencies in the country's economy.



The Chimp and the River: How AIDS Emerged from an African Forest
David Quammen W. W. NORTON *(2015)*
This intense study of the origins of AIDS is excerpted and adapted by David Quammen from his book *Spillover* (W. W. Norton, 2012; see N. Wolfe *Nature* **490,** 33; 2012). With Sherlockian verve, Quammen traces the trail from the first human cases, through labs around the world, and finally to virologist Beatrice Hahn's discovery that simian immunodeficiency virus (SIV), from which HIV-1 is derived, can kill wild chimpanzees. Quammen's portrait of the real 'Patient Zero' as a Cameroonian hunter clumsily butchering a chimp is a masterful summing-up of the evidence.

Science in Wonderland: The Scientific Fairy Tales of Victorian Britain
Melanie Keene OXFORD UNIVERSITY PRESS *(2015)*
The prodigious pace of Victorian research — from the unearthing of dinosaur fossils to the laying of a transatlantic telegraph cable — posed a stiff pedagogical challenge. To deliver the new findings on nature to the public, writers seized on the era's obsession with the supernatural. Science historian Melanie Keene argues here that many "fairy tales of science" were educational gems: by harnessing tropes of the genre to communicate facts, they evoked a scientific wonder that truly came into its own in the age of quantum mechanics and relativity. (See M. Keene *Nature* **504,** 374–375; 2013.)



Climate Shock: The Economic Consequences of a Hotter Planet
Gernot Wagner and Martin L. Weitzman PRINCETON UNIVERSITY PRESS *(2015)*
Economists Gernot Wagner and Martin Weitzman deliver a high-voltage shock in their analysis of the costs of climate change. With uncurbed emissions predicted to rise steeply by 2100, a radical reframing of the catastrophe as a global risk-management issue is due, they argue. Their blueprint is a three-step response: scream (call for business and policy-makers to snap to it); cope (adapt rapidly to events); and profit (invest in green industry). **Barbara Kiser**

KEVIN FRAYER/GETTY

China is setting records for installing solar panels — even as most of the country's energy comes from coal.

SUSTAINABILITY SCIENCE

Exploiting the synergies

Dave Griggs relishes Jeffrey Sachs's analysis of the policy and practice key to a viable future for people and planet.

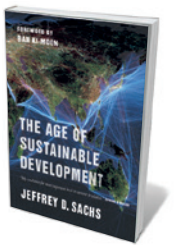
As a concept and practice, sustainable development emerged on the global scene in 1972, with the United Nations Conference on the Human Environment. Four decades on, in the year that the United Nations is due to set its Sustainable Development Goals (SDGs), the idea remains fuzzy around the edges. Jeffrey Sachs's *The Age of Sustainable Development* sharpens our understanding. It is, in my view, the best, most comprehensive and most articulate exposition of sustainable development ever written.

Sachs is a rock-star economist, leading thinker in sustainable development and senior UN adviser. *The Age of Sustainable Development* is based on his excellent massive open online course (MOOC) of the same name.

He defines sustainable development as a "normative outlook" aiming to solve global problems such as climate change through environmental, economic and social goals, along with good governance. He shows that it is a science of complex systems: the global economy, the Earth system, politics and social interactions such as support networks and social media.

NATURE.COM
For more on science in culture, see:
nature.com/booksandarts

Sustainable development was once considered a problem of developing countries, solvable through, and almost as a by-product of, economic growth. But no country has pulled itself out of poverty without fossil fuels, whose emissions drive climate change and pollution, or nitrogen-based fertilizers, which promote algal blooms. And richer countries have demonstrated the problems of uncontrolled development of land and resources, a factor in biodiversity loss. Sustainable development is crucial for all countries, so the SDGs will apply to every nation.



The Age of Sustainable Development
JEFFREY D. SACHS
Columbia Univ. Press: 2015.

Sachs recognizes the benefits of economic growth, citing the case of China, which has achieved history's most remarkable economic transformation, with extreme poverty falling from 84% in 1981 to just 12% in 2010. However, he also shows the limitations of growth through challenges still affecting billions, from poverty to food security. He explains some of how we got to where we are today, highlighting the economic and social factors that maintain the status quo or make things worse, such as the historic, geographical and political forces that are widening inequalities in countries such as the United States.

How to achieve a sustainable future? Education, Sachs notes, is a lynchpin. When girls stay in school for longer, fertility rates drop. Households with fewer children invest more in education, health and nutrition. He quotes Scottish economist Adam Smith, who wrote in *The Wealth of Nations* (1776) that because society benefits when people are educated, the costs should be "defrayed by the general contribution of the whole society". That we have not achieved this more than two centuries later is a baffling and damning indictment.

Alongside the social challenges are climate change, ocean acidification and the current mass extinction of species — serious threats to humanity's capacity to thrive or even survive. For example, the concentration of carbon dioxide in the atmosphere is rising by more than 2 parts per million each year. Sachs concludes that no country is currently on a path to sustainable development.

What becomes clear is that understanding the links between these issues is essential. Along with development aims such as sanitation and health care for a growing and ageing population, there are environmental challenges such as mitigating climate change. It is important to pinpoint solutions with positive trade-offs, such as encouraging people to walk or cycle, to reduce emissions while conferring health benefits. It is equally important to avoid fixing one problem by exacerbating another, for example providing universal access to affordable energy by burning more fossil fuels (see page 151).

Sachs struggles with this, as do businesses and governments. He compartmentalizes the book into chapters dealing with issues such as health, food security and climate change, which fails to show the interdependent nature of the beast in all its horrifying complexity. But in a finale on the SDGs, he delivers a unified message clearer, more insightful and more accessible than previous attempts.

Sachs explains the benefits of goal-based development such as mobilizing knowledge and practice networks — most importantly, those that include the scientific community, the public, politicians and non-governmental organizations. He explains how they might be financed through the public and private sector, and governed with accountability, transparency and participation.

I would make this book compulsory reading for all politicians and business leaders. ■

David Griggs *is a professor of sustainable development at Monash University in Melbourne, Australia, and the University of Warwick, UK.*
e-mail: dave.griggs@monash.edu

Correspondence

Stamp out shabby research conduct

As heads of funding bodies for medical research, we are concerned that questionable practices among researchers seem to be becoming more prevalent. Although these do not meet current definitions of misconduct, they can still distort biomedical science and cause irreproducibility — with potentially critical consequences for policies and patients.

For example, researchers may cut corners by withholding methodological details or by failing to disclose data for independent scrutiny. Inadequate training can also be responsible for false conclusions arising from flawed experimental design, methodology or statistical analysis. Some countries, including Australia, Canada and the Netherlands, have a category for these — 'poor conduct'. This must be addressed if proved, even though it is less egregious than research misconduct.

International funding bodies, informally convening with heads of international biomedical research organizations, have agreed to undertake a worldwide analysis of definitions of different types of misconduct and the policies used to tackle them. This should help to harmonize standards of research rigour and integrity globally, for the ultimate benefit of patients.

Warwick P. Anderson* *National Health and Medical Research Council of Australia, Canberra, Australia.*
warwick.anderson@nhmrc.gov.au
**On behalf of 4 correspondents (see go.nature.com/adxrpe for full list).*

Tax transactions to stabilize trading

An obvious method for controlling high-speed trading (M. Buchanan *Nature* **518,** 161–163; 2015) is a global financial-transaction tax of the kind proposed by the European Union in 2011.

Such a tax, originally designed to raise revenue, could be set to lead to a typical trading time. There would be no need to regulate trading times explicitly: it would simply not be profitable to trade on the tiny, rapid fluctuations that now trigger transactions. This solution would be simpler — and, with its revenues, more beneficial — than technical approaches such as 'speed bumps' that delay transactions.

The non-equilibrium, complex systems that correspond to economies often operate at the threshold of instability. Adding taxes ('friction') should not be seen as creating inefficiency, but as a stabilizing influence that can avoid the costs of dramatic crashes.

John Bechhoefer *Simon Fraser University, Burnaby, Canada.*
johnb@sfu.ca

Undergraduate research in action

Our programmes at California State University address the challenges of bringing undergraduates into research labs (see *Nature* **518,** 127–128; 2015). The students are then better equipped for admission to the top professional training programmes in the United States and worldwide.

More than 100 undergraduate research students are trained every year under our programmes, which have been running for 43 years and have garnered a US Presidential Award for Mentoring, among other honours. We aim to make students proficient in doing quality research experiments and in statistically analysing and publishing them.

All new undergrads are trained by peer undergraduates (not graduate students or postdocs) experienced in the research, ensuring that the newcomers immediately feel comfortable in the research setting; their work is regularly checked by senior staff.

The burden of heavy course loads is mitigated by an open-lab policy that allows students to pursue their research out of hours and during university vacations.

Our undergraduates have co-authored hundreds of publications and national presentations. And, to stimulate pre-college students' interest in research, we established a journal of student research abstracts almost 20 years ago (now open access), and annual symposia for student research posters (see go.nature.com/dwox6d).

Steven B. Oppenheimer *California State University, Northridge, California, USA.*
steven.oppenheimer@csun.edu

Women's grants lost in inequality ocean

Denmark last year launched its YDUN programme, an experimental one-year government research-funding scheme specifically for women. It was branded as sexist and provoked a political squall, so is unlikely to be repeated. Our analysis indicates that the 110 million krone (US$16 million) allocated to YDUN is roughly the same as the shortfall in Danish grant money won by women compared with men every year over the past 10 years.

The proportion of successful grant applications in 2009–13 to the Danish Council for Independent Research (DFF), which also ran YDUN, was roughly comparable for male and female researchers according to their own analysis (14% and 11%, respectively; see go.nature.com/uryhca (in Danish)).

However, our analysis of DFF data since the council's foundation in 2005 revealed that this 3% difference in success rates is significant: it corresponds to a male advantage of an average of 104 million krone per year, comparable to the entire YDUN funding allocation for women.

YDUN was a welcome attempt to widen Denmark's talent pool, but managed to level the playing field for only one year, and only for the DFF. Even then, the success rate for YDUN was only 3% (17 of 553 applicants). This level of competition is much higher than for DFF funding. Even though YDUN funding effectively made up the shortfall within the DFF for 2014, women still had to compete much harder to get it.

Darach Watson, Jens Hjorth *Niels Bohr Institute, University of Copenhagen, Denmark.*
darach@dark-cosmology.dk

Assessing resistance to new antibiotics

Losee Ling and colleagues detect no bacterial resistance to the new antibiotic molecule teixobactin (L. L. Ling *et al. Nature* **517,** 455–459; 2015), but this could be because the conditions of their test may limit its sensitivity (see J. Ramsayer *et al. Evol. Appl.* **6,** 608–616; 2013). 'Evolutionary rescue' is a more powerful assay for evaluating the probability of resistance to novel antibiotics in large bacterial samples, and therefore for informing decisions about their usage.

Evolutionary rescue assays can distinguish between resistant mutants that are present initially and those that emerge later (H. A. Orr and R. L. Unckless *PLoS Genet.* **10,** e1004551; 2014). This type of assay can also be used to evaluate factors that contribute to the emergence of bacterial resistance, such as 'horizontal' gene transfer from other bacteria or the presence of bacterial 'mutator' strains with vastly increased mutation rates.

Michael E. Hochberg *Université Montpellier, France.*
mhochber@univ-montp2.fr
Gunther Jansen *Christian-Albrechts-Universität, Kiel, Germany.*

Yves Chauvin

(1930–2015)

Nobel-prizewinning chemist who rearranged carbon–carbon bonds.

ALAIN JOCARD/AFP/GETTY

The impact of Yves Chauvin's work across the chemical industries is mind-boggling. By dissecting how carbon–carbon bonds shift in reactions of petroleum compounds, Chauvin revealed the steps in one of organic chemistry's most important reactions: metathesis. This Nobel-prizewinning work laid the path for chemical processes that are now used to make everything from pesticides to drugs. His proudest achievement, however, was developing crucial processes in the oil and plastics industries, now used to produce millions of tonnes of compounds each year.

Chauvin, who died on 27 January, was born in 1930 in Belgium, close to the border with France. His French parents sent him across the border daily to primary school; when his family returned to France, Chauvin finished his education in Paris. His summers were spent in a large family house in Tours, in France's Loire Valley, where he lived out the end of his life.

After finishing his undergraduate degree in chemical engineering in 1954 at Lyon's college of industrial chemistry (École Supérieure de Chimie Industrielle de Lyon), he began working at the French chemical company Progil (now part of Sanofi), where he met his wife, Huguette Labarre.

Chauvin said that he regretted that military service and other circumstances kept him from pursuing a PhD. But he also felt that not having one freed his mind to consider a broad range of topics. He resigned from Progil after two years because managers demanded that he simply copy procedures without exploring ideas from other fields.

In 1960, he moved to his scientific home for the next 40 years, the French Institute of Petroleum (IFP) near Paris. Here, Chauvin devoted himself to accelerating the production of chemicals by a process known as homogeneous catalysis. In this procedure, all components are dissolved in a solution, enabling fine control and the ability to work with large volumes of chemicals at relatively low temperatures. He bucked the trend in petrochemistry that then favoured catalysis on solid substrates, a technique that requires higher temperatures and often produces toxic by-products.

The work led to the invention of processes that are now central to the petrochemical industry. The dimersol and difasol processes pair smaller hydrocarbons to 'octane boosters' added to petrol or, in a modified version, to the starting material for plasticizers, additives that increase the plasticity or fluidity of a material. The alphabutol process combines carbon molecules to make feedstocks and additives for everything from lubricants to plastics.

Chauvin continued to develop homogeneous catalysis. He solved a major problem: the separation of the catalyst from the reaction medium, drawing on his knowledge of batteries to develop ionic liquids as new solvents.

Although Chauvin's name became synonymous with the process of homogeneous catalysis, he is best known for working out the steps of the intricate molecular dance known as olefin metathesis. Here, fragments of olefins — molecules containing double-bonded carbon atoms — swap places with each other, much as dancing couples swap partners. The genius of Chauvin was to co-opt ideas from a very different chemical process called ring-opening polymerization.

In a simple but ground-breaking experiment, he reacted two types of olefin (cyclic and non-cyclic) to show that the resulting products combined fragments of both. He deduced that the molecular swaps were not symmetrical, as was commonly assumed, but were orchestrated by the formation of a temporary hydrocarbon ring containing the metal catalyst. Other chemists, notably Robert Grubbs and Richard Schrock, with whom Chauvin shared the 2005 Nobel Prize in Chemistry, used this insight to improve and develop industrial reactions that underpin much of 'green' chemistry — efficient industrial processes that produce little waste.

When Chauvin retired from the IFP in 1995, he came to work in my surface organometallic chemistry laboratory at the University of Lyon. He would regularly catch the 5 a.m. train from his home in Tours to be at the bench by 8 a.m.

Chauvin's later collaborations adapted homogeneous catalysis to ionic liquids, which can be effectively applied to a variety of reactions, and their products — used for many industrial purposes — are readily isolated from the catalyst.

Yves's scientific virtuosity was tempered with humility. He was reluctant to go to Stockholm in 2005 because he felt that his contribution was less than that of Grubbs and Schrock, who made metathesis reactions broadly practical. He balanced fundamental and applied research, producing more than 100 papers and 130 patents.

Yves was always young at heart. He never missed his 16-kilometre weekly hike or failed to read the weekly edition of *Chemical Abstracts*. His deep curiosity was equalled by a knowledge and intuition that made him a fantastic inventor. Yves is deeply missed, both as a friend and as a great scientist. ■

Jean-Marie Basset *is director of the KAUST Catalysis Center at the King Abdullah University of Science and Technology in Thuwal, Saudi Arabia. He ran the laboratory at the University of Lyon in France where Yves Chauvin worked from 1996 to 2009.*
e-mail: jeanmarie.basset@kaust.edu.sa

PLANETARY SCIENCE

Enceladus' hot springs

The detection of silicon-rich particles originating from Saturn's moon Enceladus suggests that water–rock interactions are currently occurring inside it — the first evidence of ongoing hydrothermal activity beyond Earth. SEE LETTER P.207

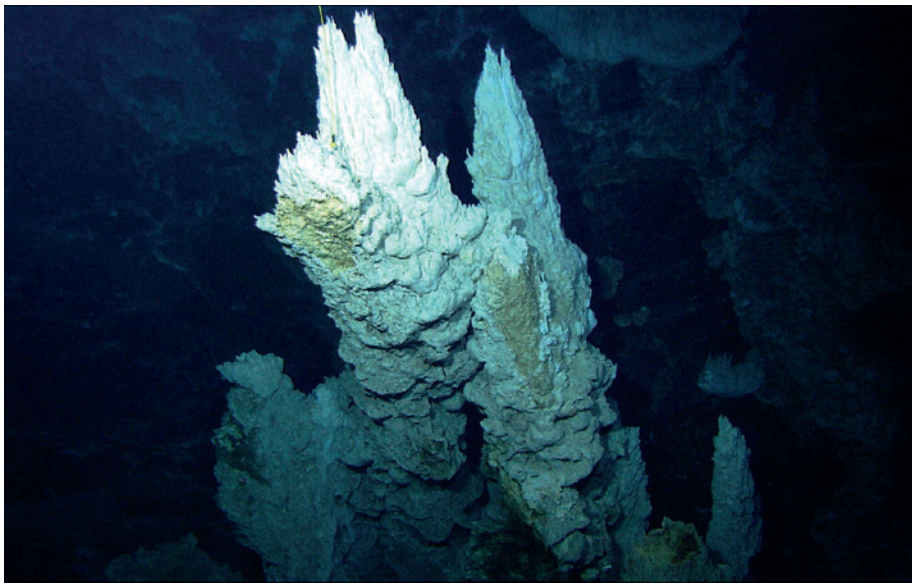
GABRIEL TOBIE

Many planetary bodies are thought to have produced hydrothermal activity — interactions between water and rock — as a result of hot-water circulation during the early stages of the Solar System, but Earth was the only one known to be sustaining such activity today. Then, in 2005, the Cassini spacecraft discovered eruptions of water vapour and ice emanating from long, warm fractures on the south pole of Saturn's moon Enceladus. The detection of salted, icy grains in Enceladus' erupting plume[1] clearly pointed to an ocean environment below its icy crust and to the leaching of rocks by warm water, at least in the past. On page 207 of this issue, Hsu *et al.*[2] report hints of presently active hydrothermal processes on Enceladus.

This story began about a decade ago during Cassini's approach to Saturn, when one of the spacecraft's instruments detected tiny dust particles, called stream particles, escaping into interplanetary space from the Saturn system[3]. Analysis of these particles revealed that they were mostly nanometre-sized and rich in silicon, in contrast to the ice-rich particles prevalent in the Saturnian environment. The origin of these particles has remained enigmatic for years.

Building on their earlier modelling work[4], Hsu and colleagues simulated the dynamics of the particles' ejection, tracking them back to their most probable source region: Saturn's E ring, a tenuous ring mostly made of small ice grains, extending between the orbits of the moons Mimas and Titan. Because Enceladus is the source of particles in the E ring[5], it must also be the ultimate source of the silicon-rich stream particles, which were presumably once incorporated in icy grains.

By analysing mass spectra of the stream particles, the authors concluded that the dominant constituent is silica (SiO_2). This is much more probable than pure silicon or silicon carbide (SiC), two other potential candidates. Silica is extremely common on Earth, occurring mostly in the natural form of quartz. But finding silica nanoparticles in the Saturnian environment is unexpected. Hsu *et al.* ruled out fragmentation of larger grains as a possible process to explain the narrow size distribution of the stream particles. The composition and size distribution must therefore be inherited from the particles' formation process, which seems most likely to have been fast crystallization of silica nanoparticles from supersaturated aqueous solutions.



IFE, URI-IAO, UW, LOST CITY SCIENCE PARTY; NOAA/OAR/OER; LOST CITY 2005 EXP./CC BY 2.0

Figure 1 | The Lost City hydrothermal field under the mid-Atlantic Ocean. These limestone chimneys, which are up to 60 metres tall, vent fluids at a temperature of 90 °C. Hsu *et al.*[2] report evidence of a similar aqueous environment in Saturn's icy moon Enceladus.

Using laboratory experiments, the authors finally showed that silica particles with the observed size distribution can be produced only under rather specific thermo-physical conditions, thus constraining the thermal state of Enceladus' interior. Specifically, a region of the rock core must have a temperature of at least 90 °C and be in contact with water of pH greater than 8.5 to dissolve silica in sufficient amounts; the oceanic salinity should be less than 4% and oceanic pH in the range of 8.5 to 10.5, to allow the formation of numerous nanometre-sized silica grains.

The inferred core temperature is unexpectedly high for a body the size of Enceladus (approximately 500 kilometres in diameter), especially given that deep water circulation should efficiently cool the core[6]. A strong heat source must exist to raise the core temperature above 90 °C — most probably tidal friction, and possibly also exothermic water–rock reactions known as serpentinization reactions[7]. But modelling is needed to determine whether tidal flow and serpentinization in the core could provide sufficient energy at present to allow hydrothermal activity, and, if so, for how long.

Intriguingly, the conditions inferred by Hsu and colleagues in Enceladus' water–rock system are similar to those found on Earth in an atypical hydrothermal field called Lost City (Fig. 1), which was discovered in the early 2000s in the mid-Atlantic Ocean[8,9]. This hydrothermal field consists of limestone chimneys 60 metres tall, which vent metal-poor, basic fluids (pH 10–11) at a temperature of 90 °C; the fluids are rich in hydrogen, abiotically produced methane and other organic compounds. For comparison, most other known fields are fuelled by acidic (pH 3–5), metal- and sulfide-rich fluids at temperatures greater than 300 °C (ref. 8).

Because it is relatively cold, Lost City has been posited[9] as a potential analogue of hydrothermal systems in active icy moons. The current findings confirm this. What is more, alkaline hydrothermal vents might have been the birthplace of the first living organisms on

the early Earth, and so the discovery of similar environments on Enceladus opens fresh perspectives on the search for life elsewhere in the Solar System.

Hsu *et al.* also conclude that the silica particles must be transported from the core hydrothermal source to the plume source near the surface in a fairly short time — from months to years at most — to limit the particles' growth. This implies that samples of materials erupted from Enceladus' warm fissures would provide a unique opportunity to directly probe aqueous, possibly prebiotic, processes occurring deep in Enceladus' rock core, in almost real time. Cassini's discoveries, together with Hsu and colleagues' findings, point to potentially complex chemical processes in Enceladus' watery interior. Cassini will fly through the moon's plume again later this year, but only future missions that can undertake improved *in situ* investigations[10,11], and possibly even return samples to Earth[11], will be able to confirm Enceladus' astrobiological potential and fully reveal the secrets of its hot springs. ■

Gabriel Tobie *is at the Laboratoire de Planétologie et Géodynamique, Université de Nantes, CNRS, UMR-6112, Nantes, France.*
e-mail: gabriel.tobie@univ-nantes.fr

1. Postberg, F., Schmidt, J., Hillier, J., Kempf, S. & Srama, R. *Nature* **474,** 620–622 (2011).
2. Hsu, H.-W. *et al. Nature* **519,** 207–210 (2015).
3. Kempf, S. *et al. Science* **307,** 1274–1276 (2005).
4. Hsu, H.-W. *et al. J. Geophys. Res. Space Phys.* **116,** A09215 (2011).
5. Kempf, S. *et al. Icarus* **193,** 420–437 (2008).
6. Travis, B. J. & Schubert, G. *Icarus* **250,** 32–42 (2015).
7. Malamud, U. & Prialnik, D. *Icarus* **225,** 763–774 (2013).
8. Kelley, D. S. *et al. Nature* **412,** 145–149 (2001).
9. Kelley, D. S. *Oceanography* **18,** 32–45 (2005).
10. Tobie, G. *et al. Planet. Space Sci.* **104,** 59–77 (2014).
11. McKay, C. P., Anbar, A. D., Porco, C. & Tsou, P. *Astrobiology* **14,** 352–355 (2014).

CELL SIGNALLING

Disarming Wnt

The secreted enzyme Notum has been found to inhibit the Wnt signalling pathway through removal of a lipid that is linked to the Wnt protein and that is required for activation of Wnt receptor proteins. SEE ARTICLE P.187

ROEL NUSSE

Cells signal to one other through secreted molecules that are conserved across the evolutionary spectrum. One class of these signals is Wnt proteins, which influence the balance between proliferation and differentiation in many cell types, including stem cells[1]. Because this balance is crucial for normal tissue maintenance, and overactivation of Wnt signalling can cause cancer, the activity of Wnt signals is tightly controlled by various extracellular molecules. In this issue, Kakugawa *et al.*[2] (page 187) describe an unexpected mechanism by which Wnt signals can be downregulated, showing how an extracellular enzyme called Notum renders Wnt inactive.

Detailed biochemical, structural and genetic experiments[1] reveal that Wnt signalling mechanisms are built from unusual elements. When Wnt proteins are made, an acyl group from palmitoleic acid (a monounsaturated form of the lipid palmitic acid) is attached at an evolutionarily conserved serine amino-acid residue, through a carboxyl ester link[3,4]. This modification is made by Porcupine, an enzyme located in a cellular substructure called the endoplasmic reticulum[5]. Such palmitoleoylation is essential for Wnt activity, because it is the acyl group that binds to Frizzled[6] — the transmembrane receptor protein for Wnt — through a hydrophobic cavity in the receptor on target cells. Wnt–Frizzled binding is imperative for receptor activation, and triggers many events in the cell, from modulating gene expression to changing cell shape.

As with many components of the Wnt signalling pathway, Notum was originally discovered in fruit flies, in screens for genes that interact with the Wnt protein Wingless[7,8]. Loss of Notum in flies leads to abnormal wing growth, indicating that Wnt signalling (which drives wing growth and patterning) becomes unrestrained in its absence. Wnt signals also turn on the expression of the gene that encodes Notum, leading to negative feedback regulation that intrinsically limits signalling, as is often the case for such pathways. Initial studies[7,9] suggested that Notum might act as a phospholipase enzyme, cleaving the link between membrane-bound glycoproteins called GPI anchors and glypicans — large polysaccharides that form complexes with extracellular molecules such as Wnt. Cleavage releases glypicans into the extracellular space, decreasing their ability to restrain Wnt activity.

Kakugawa *et al.*[2] unveil a different, previously undocumented function of Notum. The authors start with a structural analysis of human and fly Notum, and find that the protein has the overall structure of a hydrolase enzyme. But it also has a large hydrophobic cavity of around 380 cubic ångströms, which in theory could provide sufficient space for binding by acyl groups with chains of up to 16 carbons — the length of palmitoleic acid. Furthermore, the researchers' analysis of binding between Notum and acyl groups of various lengths and

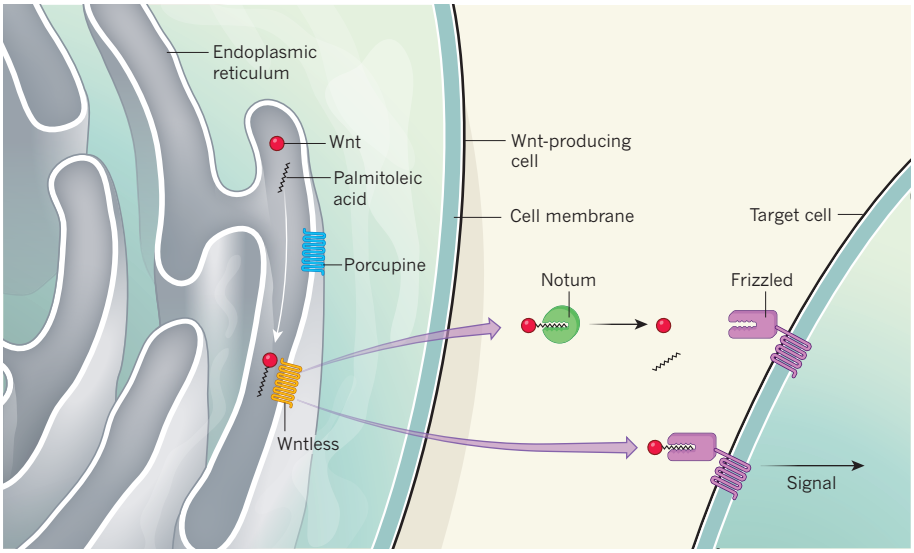


Figure 1 | Notum shoots the messenger in Wnt signalling. In Wnt-producing cells, the Wnt protein is made in a cellular compartment called the endoplasmic reticulum. There, an acyl group from palmitoleic acid is added to Wnt by the membrane-spanning enzyme Porcupine. The Wntless protein then transports palmitoleoylated Wnt out of the cell. Secreted Wnt binds to its receptor protein Frizzled, which spans the membrane of Wnt target cells. This binding depends on the acyl group in Wnt, and triggers an intracellular signalling cascade. Kakugawa *et al.*[2] report that the Wnt–Frizzled interaction is inhibited by the extracellular enzyme Notum, which specifically removes the acyl group from Wnt.

degrees of saturation reveals that, of the longer-chain molecules, only monounsaturated molecules can bind. In other words, the acyl group found on Wnt can form a complex with Notum, whereas lipids with different configurations cannot. In parallel with their binding assays, Kakugawa and colleagues show that Notum enzymatically removes the acyl group from Wnt, thereby rendering the protein inactive. Such an extracellular deacylase activity has never been previously reported.

Hedgehog is another signalling molecule whose activity is modified by lipids. But the authors demonstrate that, unlike Wnt, Hedgehog is not a substrate for Notum. The specificity of Notum for monounsaturated acyl groups provides an explanation for this discrepancy, because the acyl group attached to Hedgehog contains saturated carbon bonds throughout. Kakugawa and co-workers also provide genetic evidence that, in flies, Notum does not interact with Hedgehog signalling *in vivo*. Finally, they show that Notum contains binding sites for polysaccharides such as glypican sugar chains, inviting speculation that glypicans bring together Notum and Wnt — thus modulating the enzymatic interaction of Notum with Wnt, rather than acting as a substrate for Notum to cleave GPI anchors.

Kakugawa and colleagues' discovery adds greatly to our understanding of Wnt signalling, and of the central role of the Wnt lipid group. The authors' results demonstrate how acquisition or loss of the acyl group from palmitoleic acid can adroitly control the activation or deactivation of Wnt signals. The transmembrane protein Wntless conveys Wnt molecules that have been palmitoleoylated by Porcupine through the cell for secretion[10]. Once secreted, Wnt proteins bind to Frizzled on other cells through the acyl group (Fig. 1).

All of these lipid-related pathway components, including Notum, evolved at around the same time as Wnt. The Wnt protein itself contains a lipid-binding motif called a saposin fold, and it has been speculated[11] that, when Wnt signals initially evolved, they consisted of a lipid–protein complex, with the two becoming covalently linked at a later date. Lipids lie at the heart of Wnt signalling, and can even be viewed as a primordial cell-fate signal because they are also used by organisms such as choanoflagellates, which are located at the base of the animal evolutionary tree[12].

Because enzymes are often good targets for drugs, it might be possible to identify molecules that inhibit the activity of Notum, thereby increasing the strength of Wnt signalling. Wnt proteins can stimulate stem cells to proliferate, so such an approach could have therapeutic value for treating degenerative diseases. Collectively, these findings explain how Notum prevents tissues from growing abnormally or adopting aberrant identities — it shoots the messenger in the Wnt pathway by stripping Wnt proteins of their crucial lipid group. ■

Roel Nusse *is at the Howard Hughes Medical Institute, Department of Developmental Biology, Institute for Stem Cell Biology and Regenerative Medicine, Stanford University School of Medicine, Stanford, California 94305-5458, USA.*
e-mail: rnusse@stanford.edu

1. Clevers, H., Loh, K. M. & Nusse, R. *Science* http://dx.doi.org/10.1126/science.1248012 (2014).
2. Kakugawa, S. *et al. Nature* **519,** 187–192 (2015).
3. Willert, K. *et al. Nature* **423,** 448–452 (2003).
4. Takada, R. *et al. Dev. Cell* **11,** 791–801 (2006).
5. Rios-Esteves, J., Haugen, B. & Resh, M. D. *J. Biol. Chem.* **289,** 17009–17019 (2014).
6. Janda, C. Y., Waghray, D., Levin, A. M., Thomas, C. & Garcia, K. C. *Science* **337,** 59–64 (2012).
7. Giráldez, A. J., Copley, R. R. & Cohen, S. M. *Dev. Cell* **2,** 667–676 (2002).
8. Gerlitz, O. & Basler, K. *Genes Dev.* **16,** 1055–1059 (2002).
9. Häcker, U., Nybakken, K. & Perrimon, N. *Nature Rev. Mol. Cell Biol.* **6,** 530–541 (2005).
10. Coombs, G. S. *et al. J. Cell Sci.* **123,** 3357–3367 (2010).
11. Bazan, J. F., Janda, C. Y. & Garcia, K. C. *Dev. Cell* **23,** 227–232 (2012).
12. Alegado, R. A. *et al. eLife* **1,** e00013 (2012).

This article was published online on 25 February 2015.

EVOLUTION

Fitness tracking for adapting populations

A method for tracking the descendants of hundreds of thousands of yeast cells in an evolving population reveals that thousands of individuals contribute to early increases in population-wide fitness. **SEE ARTICLE P.181**

DAVID GRESHAM

Positive selection for genetic variants that benefit an organism in a particular environment, a process called adaptive evolution, affects all species. As such, knowing how frequently beneficial mutations occur, and quantifying the selective advantage they confer — their fitness — has been a long-standing goal for evolutionary biology[1]. On page 181 of this issue, Levy *et al.*[2] describe a method for tracking individual genetic variants in an evolving population, and measuring their fitness and fate as the population adapts to the environment.

Individuals descended from a common progenitor are said to be of the same genetic lineage. Levy and colleagues tracked individual lineages in yeast (*Saccharomyces cerevisiae*) with extremely high resolution by introducing hundreds of thousands of unique, random DNA sequences into individual yeast cells that have otherwise-identical genomes. These sequences, called barcodes, have no impact on the cell, but can be used to distinguish between different individuals by means of DNA sequencing. Individuals that have the same barcode are part of the same lineage, allowing estimation of how many cells in the population are descended from a common ancestor.

After barcoding the yeast cells, the authors studied the population as it underwent adaptive evolution over many generations in a simple environment. In the evolving population, each daughter cell is born through cell division and so is a clone of its mother. Thus, sexual reproduction plays no part in the population's evolutionary dynamics. Although all cells in the population start out with identical genomes (apart from the barcodes), genetic diversity is introduced by random mutations that arise spontaneously when DNA is replicated during cell division. If a mutation is beneficial in the environment, allowing the cell

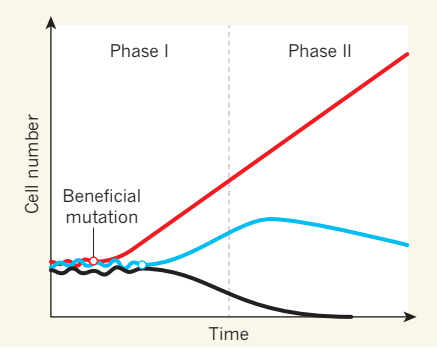


Figure 1 | Get fit or die trying. Levy *et al.*[2] labelled hundreds of thousands of individual yeast cells, and tracked the population as it evolved. Each lineage is initially present in approximately equal numbers. In the first, predictable phase of evolution, thousands of lineages that acquire beneficial mutations (blue, red) expand, increasing the fitness of the population and leading to the decline of lineages that did not acquire beneficial mutations (black). In a second, less predictable phase of evolution, even lineages with beneficial mutations can decline (blue), as those containing mutations that confer an exceptionally high degree of fitness, and that arose early enough, continue to expand (red), further increasing population fitness. (Adapted from Fig. 1a of ref. 2.)

and its descendants to proliferate more rapidly, that lineage will begin to increase in relative abundance in the population. By sequencing the cells' molecular barcodes at different time points throughout the experiment, beneficial lineages can be identified.

Levy and colleagues used their high-resolution lineage-tracking technique to quantify the fitness of each beneficial lineage, and to determine when the corresponding mutation occurred in the population's history. They found that, in evolving yeast populations containing 70 million cells, about 25,000 lineages showed fitness increases of more than 2% after just over 100 generations. Many of these lineages were present at frequencies lower than 0.001%. This means that there are initially many more competing lineages containing beneficial mutations in evolving populations than previously revealed by whole-population sequencing[3–5].

The aggregate effect of these thousands of beneficial lineages is to push the population fitness higher and higher. In doing so, a process of sequential purging occurs. First, the lineages that did not acquire a beneficial mutation are removed from the population. Then, as population fitness continues to increase, even lineages that contain beneficial mutations are purged once their individual fitness is less than that of the population as a whole.

Levy and colleagues' study shows that there are two distinct phases in the adaptive evolution of a large cell population (Fig. 1). In the first phase, population fitness increases in a predictable manner. This increase is attributable to the cohort of thousands of different lineages with beneficial mutations, and depends on the size of the population and the fitness associated with each mutation. The second phase is less predictable. The ultimate 'winners' must have higher fitness than the overall population and the mutations must have been introduced early enough in the population's history to establish themselves — this phase is unpredictable because such mutations are rare.

The ability to quantify the fitness of each beneficial mutation in a population enabled Levy and co-workers to measure the range of fitnesses conferred by beneficial mutations. Theory predicts[6,7] that the distribution of fitness effects associated with new mutations has a particular mathematical shape, known as an exponential distribution. However, the authors find that this is not the case, at least not in this environment. Instead, they observe a complicated distribution of fitness effects that seems to be composed of a mixture of distributions, which may reflect beneficial mutations in different genes. The nature of the distribution of fitness effects of beneficial mutations is central to understanding and simulating adaptive evolution in future experiments. As such, the ability to empirically measure this distribution with precision provides opportunities to reconcile theory and data.

Despite the power of Levy and co-workers technique, several limitations remain. First, the method does not actually identify the beneficial mutations, a key requirement for understanding the molecular basis of adaptation[8,9]. Second, it tells us about the distribution of fitness effects for beneficial mutations, which are most relevant to the evolution of large asexual populations, but not those for neutral or deleterious mutations, which may be important in populations that are small, sexual or have high mutation rates. Last, and crucially, the method in its current form allows identification of only the earliest stages of adaptive evolution. Once a single lineage has swept to high frequency in the population, its barcode will be abundant. Loss of barcode diversity limits the ability to detect a second beneficial mutation within these lineages, a problem that could be overcome by somehow regenerating the diversity of barcodes during the course of the experiment.

The ability to track hundreds of thousands of individual lineages in a population is an exciting tool that allows us to address many questions in adaptive evolution. Levy *et al.* performed their experiment using enormous populations, ensuring an ample supply of mutations. However, studying the dynamics of adaptation in much smaller populations would also be informative, and will probably result in less-predictable outcomes in the early stages of adaptation. Furthermore, studying adaptation in different environments and different genetic backgrounds will be crucial for assessing the generality of the results. Application of high-resolution lineage tracking in other organisms may be useful for understanding the evolutionary dynamics of antibiotic resistance in pathogens and the evolution of human tumours. The ability to observe evolution in action with high resolution is certain to reveal unanticipated features of the universal force of adaptive evolution. ■

David Gresham *is at the Center for Genomics and Systems Biology, Department of Biology, New York University, New York, New York 10003, USA.*
e-mail: dgresham@nyu.edu

1. Eyre-Walker, A. & Keightley, P. D. *Nature Rev. Genet.* **8,** 610–618 (2007).
2. Levy, S. F. *et al. Nature* **519,** 181–186 (2015).
3. Lang, G. I. *et al. Nature* **500,** 571–574 (2013).
4. Hong, J. & Gresham, D. *PLoS Genet.* **10,** e1004041 (2014).
5. Kvitek, D. J. & Sherlock, G. *PLoS Genet.* **9,** e1003972 (2013).
6. Gillespie, J. H. *Evolution* **38,** 1116–1129 (1984).
7. Orr, H. A. *Genetics* **163,** 1519–1526 (2003).
8. Dean, A. M. & Thornton, J. W. *Nature Rev. Genet.* **8,** 675–688 (2007).
9. Gresham, D. & Hong, J. *FEMS Microbiol. Rev.* http://dx.doi.org/10.1111/1574-6976.12082 (2014).

This article was published online on 25 February 2015.

50 Years Ago

The completion of the *Flora URSS* is a scientific event of great significance not only to botanists of the Soviet Union … During the War work was almost entirely suspended as most of the authors were evacuated from Leningrad. However, incredible efforts were made to continue the work. Thus, late in the autumn of 1941 in the besieged city … an attempt was made to print Volume 11. B. A. Tikhomirov … obtained the necessary amount of paper and … this volume was printed. N. F. Goncharov, already desperately weakened by starvation, proceeded with the account of the genus *Astragalus* which made up Volume 12. Later that winter this account was defended as his thesis for the degree of doctor of biology, and in February 1942 Goncharov died of hunger … Thus, thirty-three years of work and the participation of about a hundred authors were required for the completion of … a *Flora* of 30 volumes. We remember all our colleagues, many of them long dead, who contributed to its achievement. We have done what we could. We welcome the young botanists and wish them success.
From *Nature* 13 March 1965

100 Years Ago

Insects Injurious to the Household and Annoying to Man. By Prof. G. W. Herrick; *The House-Fly,* Musca domestica, *Linn. Its Structure, Habits, Development, Relation to Disease and Control.* By Dr. C. G. Hewitt — In addition to insects in the zoological sense of the term, spiders, mites, ticks, solpugids, scorpions, and centipedes are passed in review, and the British reader cannot but feel that some compensation for not being an American is afforded by the comparatively scanty house-fauna of his native land.
From *Nature* 11 March 1915

MICROBIOLOGY

How bacteria get spacers from invaders

Bacteria use CRISPR–Cas systems to develop immunity to viruses. Details of how these systems select viral DNA fragments and integrate them into bacterial DNA to create a memory of invaders have now been reported. SEE ARTICLES P.193 & P.199

IDO YOSEF & UDI QIMRON

Less than a decade ago, immunological memory was regarded as a feature unique to vertebrates — scientists ridiculed the idea that bacteria might be able to 'remember' viruses that attack them. Yet the almost inconceivable concept of bacterial immunological memory has since been shown to exist after all[1–3]. Two papers in this issue, by Heler *et al.*[4] (page 199) and Nuñez *et al.*[5] (page 193), report major advances in our understanding of the molecular mechanism of this phenomenon.

Bacteria remember their viral invaders by sampling short DNA sequences known as protospacers from the viruses' genetic material. These sequences become integrated into the bacterium's own DNA, specifically into an array of repeat sequences called clustered regularly interspaced short palindromic repeats (CRISPRs; Fig. 1); the integrated sequences are called spacers. When a bacterium is subsequently attacked by a recognized virus, the spacers are transcribed from the array and used to guide a complex containing CRISPR-associated (Cas) proteins, which cleave protospacers in viral nucleic-acid molecules[1].

Accidental destruction of the CRISPR array could occur if transcribed spacers guide Cas proteins to cleave it, leading to catastrophic degradation of bacterial genetic material. To prevent this potential autoimmunity, some bacteria have CRISPR–Cas systems that cleave DNA targets only if they are flanked by sequences known as protospacer adjacent motifs (PAMs)[6]. The repeat sequences that flank spacers in such CRISPR arrays lack PAMs and therefore cannot be cleaved (Fig. 1).

The mechanism by which spacers are chosen so that they target only PAM-associated protospacers has remained elusive. Heler and colleagues show that the Cas9 protein in two species of *Streptococcus* bacterium selects for spacers that have the correct PAM. Before now, the protein's main known role was cleaving targeted DNA.

When the authors exchanged Cas9 proteins for others that had different PAM specificities, they found that the PAM sequence of the acquired spacers changed accordingly. These CRISPR–Cas systems therefore efficiently use Cas9's ability to recognize PAM sequences for memory as well as for cleavage, instead of having dedicated memorizing proteins develop PAM recognition from scratch. In other types of CRISPR–Cas system, such as that found in *Escherichia coli*, the memorizing proteins have an intrinsic ability to select, at least partially, for PAM-encoding protospacers through an as-yet-unknown mechanism[3].

Heler and co-workers went on to show that Cas9 is required not only for determining the PAM sequence of the acquired spacers, but also for integrating spacers into CRISPRs. This feature is peculiar to Cas9 — the proteins that cleave nucleic acids in other CRISPR–Cas systems studied are not required for integration[3], but may enhance it under certain conditions[7].

These findings add crucial details to the mechanism of molecular memorization revealed by *in vivo* studies[3,7–9]. Each spacer is integrated into the CRISPR array with a new repeat; it is known that newly integrated repeats maintain the sequence of the existing repeat on the other side of the new spacer[3], and that the integration of new spacers probably occurs by separation of the two DNA strands of this repeat[9]. But more information is needed, and an *in vitro* system that allows further mechanistic details to be uncovered has long been awaited.

Nuñez and colleagues have established just such a system: it is composed of *E. coli* memorizing proteins, a supercoiled plasmid DNA as the spacer-acceptor molecule and a double-stranded (ds) DNA that serves as a spacer-donor molecule. The researchers first demonstrated the validity of their system by using it to corroborate many of the *in vivo* characteristics of the CRISPR memorization process. They went on to analyse high-throughput sequencing of spacers inserted *in vitro*, and show that the memorizing proteins integrate spacers in the correct orientation by recognizing a specific nucleotide base in the PAM.

Importantly, their system allows the spacer donor to be easily replaced with DNA that has different sequences, end modifications and strand compositions, and thus enables the influence of these features on spacer integration to be studied. In this way, Nuñez *et al.* show that hydroxyl (OH) groups at the 3′ ends of dsDNA substrates are essential for integration. On the basis of this requirement, and of characterization of intermediates identified *in vivo*[9], the authors propose a highly plausible model for spacer insertion. In this model, the memorizing enzymes catalyse bond formation between the 3′ end of a preferred strand of the spacer and a particular strand at the end of a repeat. This is followed by the formation of another bond between the 3′ end of

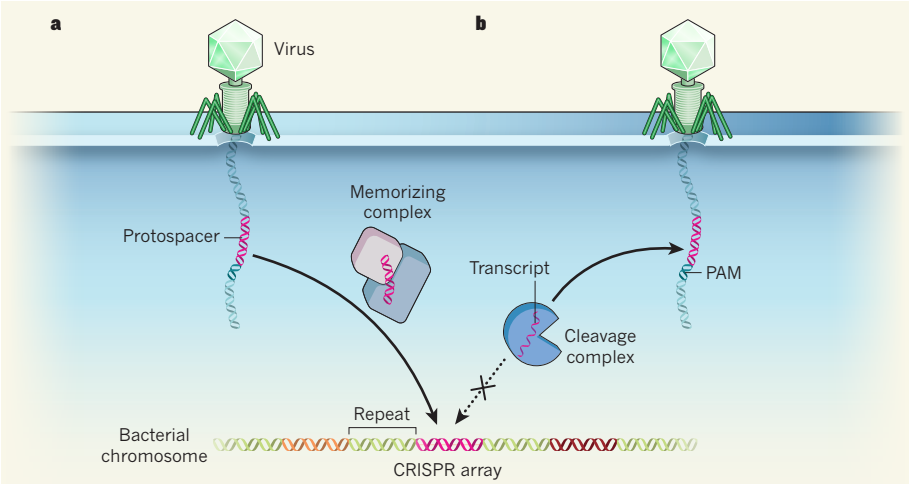


Figure 1 | The bacterial immune response. **a**, When bacteria are invaded by viral DNA, memorizing protein complexes of the CRISPR–Cas immune system select short sequences (protospacers) of the foreign DNA and integrate them as spacer sequences into their own chromosome. The spacers are integrated into an array of repeat sequences called clustered regularly interspaced short palindromic repeats (CRISPRs); different spacers are shown in orange, pink and red. **b**, If the bacterium is subsequently exposed to previously encountered DNA, a transcript of the spacer guides a cleavage protein complex to cut out the protospacer. In some types of CRISPR–Cas system, cleavage occurs only if the protospacer is flanked by a protospacer adjacent motif (PAM), and the CRISPR array is not cleaved because the spacers lack PAMs. Heler *et al.*[4] and Nuñez *et al.*[5] report details of the molecular mechanisms by which CRISPR–Cas systems select and integrate spacers into bacterial DNA.

the complementary strand of the spacer and the complementary strand at the other end of the repeat (see Fig. 5 of the paper).

The strength of *in vitro* approaches to studying biological systems is that all the components are artificially added to the reaction; the requirements and features of each component can therefore be defined and manipulated. But differences from physiological activity may occur, stemming either from the use of a different chemical environment from that found *in vivo* or from the absence of regulatory elements. Such elements may not be essential for the generation of the end product of a reaction, but might have a key role in the physiological process.

Nuñez and co-workers report just such a difference. *In vivo* studies have revealed that spacer integration occurs predominantly at the first repeat of the CRISPR array[3,7]. By contrast, the authors observe that spacer insertion in their *in vitro* system is also distributed near other repeats, and even outside the CRISPR array. The researchers suggest that this might represent a physiological way of generating new arrays. This is a valid possibility, but regulatory elements *in vivo* or in physiological conditions probably often restrict this distribution and direct integration in a specific location.

The authors also report that PAM-encoding spacer donors are not preferred substrates for integration *in vitro*, as opposed to what has been seen *in vivo*[3]. Moreover, they observe that the length of integrated spacers may vary substantially, whereas spacers in naturally occurring arrays have a strictly defined length. These differences might be explained by the fact that the *in vitro* system simulates only the last stage of spacer integration; earlier steps in the natural process probably account for the PAM preference and for defined spacer lengths observed *in vivo*.

The differences in the *in vivo* and *in vitro* studies nevertheless highlight the cardinal question of what determines the constant length of newly acquired spacers *in vivo*. Is it dictated by a protein complex that hands the processed spacer to the memorizing enzymes? If so, then what are these proteins? An *in vitro* system composed of all of the elements that catalyse every step of the reaction is needed to address these issues.

PAMs prevent autoimmunity against the CRISPR array, but autoimmunity could also occur if the CRISPR–Cas system accidentally cleaves other DNA sequences. So how is this prevented? It is known[3,10] that foreign DNA molecules are sampled by CRISPR–Cas systems more frequently than the host's chromosome. Future work should investigate the mechanism underlying this selective sampling. ■

Ido Yosef *and* **Udi Qimron** *are in the Department of Clinical Microbiology and Immunology, Sackler School of Medicine, Tel Aviv University, Tel Aviv 69978, Israel.*
e-mail: ehudq@post.tau.ac.il

1. Barrangou, R. *et al. Science* **315,** 1709–1712 (2007).
2. Brouns, S. J. J. *et al. Science* **321,** 960–964 (2008).
3. Yosef, I., Goren, M. G. & Qimron, U. *Nucleic Acids Res.* **40,** 5569–5576 (2012).
4. Heler, R. *et al. Nature* **519,** 199–202 (2015).
5. Nuñez, J. K., Lee, A. S. Y., Engelman, A. & Doudna, J. A. *Nature* **519,** 193–198 (2015).
6. Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J. & Almendros, C. *Microbiology* **155,** 733–740 (2009).
7. Datsenko, K. A. *et al. Nature Commun.* **3,** 945; http://dx.doi.org/10.1038/ncomms1937 (2012).
8. Swarts, D. C., Mosterd, C., van Passel, M. W. & Brouns, S. J. J. *PLoS ONE* **7,** e35888 (2012).
9. Arslan, Z., Hermanns, V., Wurm, R., Wagner, R. & Pul, U. *Nucleic Acids Res.* **42,** 7884–7893 (2014).
10. Nuñez, J. K. *et al. Nature Struct. Mol. Biol.* **21,** 528–534 (2014).

This article was published online on 18 February 2015.

CLIMATE CHANGE

Black carbon and atmospheric feedbacks

Climate simulations show that interactions between particles of black carbon and convective and cloud processes in the atmosphere must be considered when assessing the full climatic effects of these light-absorbing particulates.

BEN BOOTH & NICOLAS BELLOUIN

Black carbon[1], often referred to as soot, is emitted during the incomplete combustion of fossil fuels, biofuels or wood. In contrast to other particulates emitted into the atmosphere by human activities, black carbon absorbs sunlight efficiently. This absorption leads to local heating of the atmosphere, warming the planet. Black carbon has received particular interest recently[2] in the context of changes in climate policy. It remains in the atmosphere for only a few days, so cutting black-carbon emissions may be a viable way to reduce global warming over the next few decades, alongside measures to mitigate

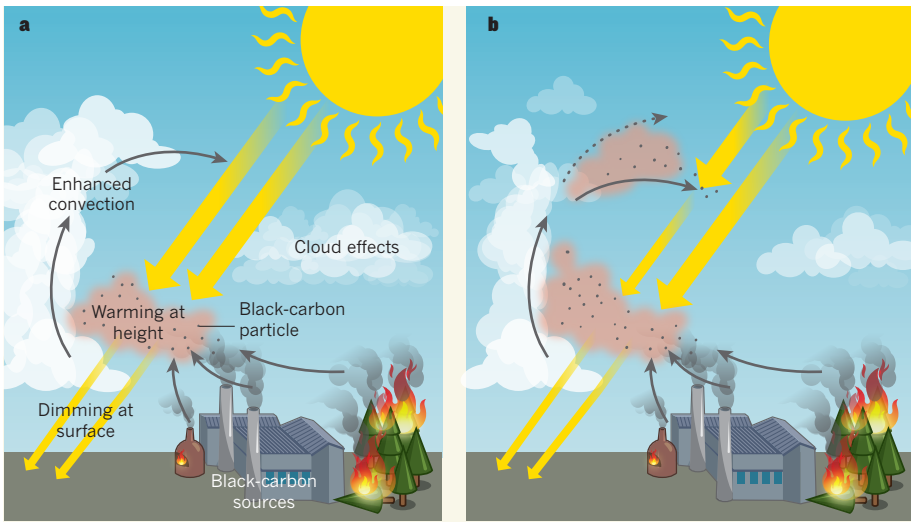


Figure 1 | A typical black-carbon feedback loop. **a**, Black-carbon emissions from sources such as industrial processes, brick kilns and forest fires have numerous influences on the atmosphere. The details of these influences can be strongly dependent on time and location, but the emissions generally lead to a net surface dimming (thin yellow arrows) alongside enhanced warming at height (pink shading). The latter factor is often associated with enhanced vertical convection and effects on clouds that have a net result, according to Sand and colleagues' climate simulations[3], of reducing cloud at altitude (not shown). **b**, Sand *et al.* suggest that roughly half of the total climate impact of black-carbon emissions is apparent only if these atmospheric responses feed back to the distribution of black carbon, lofting it to height and increasing its atmospheric lifetime and spatial extent. These changes enhance any surface dimming (extension of thin yellow arrow), and lead to greater warming at height (upper pink shading), black-carbon transport into the upper atmosphere (dotted black arrow) and further changes to clouds. Although, locally, greater warming can generate some deepened convective clouds, the net impact is a further reduction of cloud at altitude.

changes in the emissions of carbon dioxide. Such actions need to be supported by a good understanding and prediction of the climate role of black carbon. Writing in the *Journal of Climate*, Sand *et al.*[3] report climate simulations that provide insights into these issues. The results highlight challenges for upcoming international initiatives aimed at better understanding how the climate responds to changes in the composition of the atmosphere.

Sand and colleagues used a numerical simulator of Earth's atmosphere and ocean to compare the effects on climate of artificially large increases in the emission of carbon dioxide and black carbon. They find that, although the increases were designed to exert similar perturbations in Earth's energy budget (the net flow of incoming and outgoing energy), changes in the planet's surface temperature and rainfall are considerably weaker in the simulation with elevated concentrations of black carbon.

This result confirms the importance of rapid responses in the atmosphere to changes in black carbon. These responses manifest themselves as warming at height and changes in cloud properties that lead to a net decrease in mid- and high-level cloud (Fig. 1). Moreover, they act to offset the initial artificially large perturbation, mainly because the warming and cloud loss at altitude effectively radiate energy to space, before the surface climate is able to respond. However, the magnitude of the rapid responses reported by Sand *et al.* — roughly seven times stronger than those to carbon dioxide — will come as a surprise to many climate scientists.

The researchers also highlight another result, which has implications for numerical simulations of climate change. By using a pair of experiments, both of which explore the climate impacts of black carbon and differ only in whether black-carbon changes can also adjust to atmospheric-circulation responses, Sand *et al.* demonstrate the role of the two-way black carbon–atmosphere interactions in driving the full climate response. Their findings are unexpected because these interactions seem to be the dominant cause of the climate response to changes in black carbon. The change in global surface temperature varies by a factor of two between the two experiments, with considerably larger differences at altitude. Indeed, many rainfall responses appear only when feedbacks of black carbon-to-atmosphere-to-black carbon are included.

The authors point out that the feedback loop of black carbon to itself through changes in climate may be particularly strong in their simulations because their model contains an unusually active atmospheric convection. Moreover, this strength may be exacerbated further by the artificially large perturbation imposed. Experiments with other numerical models may find weaker responses. Nevertheless, the large differences in the climate impacts of black carbon, when its two-way interaction with meteorology is also included, may make it harder to determine black carbon's full climate impact.

The various groups of climate scientists each focus on specific aspects of the climate system to better understand the effects of atmospheric changes. For those who work on atmospheric particulates, such as black carbon, an important aim is to quantify the particulates' impact on Earth's energy budget. A largely separate community studies atmospheric feedbacks, such as convection and clouds. Plans are already under way to design climate-model experiments under the Coupled Model Intercomparison Project, Phase 6 (ref. 4), which will provide improved knowledge of future climate responses and feed results to the next assessment report of the Intergovernmental Panel on Climate Change. Contributions to several of these experiments will either prescribe a fixed meteorology to explore the impacts on Earth's energy balance, or use fixed concentrations of atmospheric particulates to explore atmospheric feedbacks.

Such a pragmatic approach enables groups to concentrate resources on particular aspects of the climate-change problem and gain better insight into the processes involved. However, Sand and co-workers' findings suggest that when it comes to understanding the full climate impact of black carbon, it will be crucial to account for both how black carbon influences atmospheric circulation and also how these changes feed back on the atmospheric distribution of black carbon. This highlights the risk of simplified or idealized approaches, which may produce misleading conclusions about the total climate impact of changes in black-carbon concentrations. ■

Ben Booth *is in the Met Office Hadley Centre, Exeter EX1 3PB, UK.* **Nicolas Bellouin** *is in the Department of Meteorology, University of Reading, Reading RG6 6BB, UK.*
e-mails: ben.booth@metoffice.gov.uk; n.bellouin@reading.ac.uk

1. Bond, T. C. *et al. J. Geophys. Res. Atmos.* **118,** 5380–5552 (2013).
2. Shindell, D. *et al. Science* **335,** 183–189 (2012).
3. Sand, M. *et al. J. Clim.* http://dx.doi.org/10.1175/JCLI-D-14-00050.1 (2015).
4. www.wcrp-climate.org/wgcm-cmip/wgcm-cmip6

EVOLUTIONARY BIOLOGY

The origin of terrestrial hearing

A study of the African lungfish reveals that it has a rudimentary ability to detect pressure waves caused by sound. The finding expands our knowledge of how hearing evolved in early tetrapods, the first vertebrates to have limbs and digits.

JENNIFER A. CLACK

A long-standing problem in the evolution of land vertebrates has been how they evolved to detect sound. Lungfishes are the closest living relatives of tetrapods (vertebrates that have limbs and digits), and so may help to provide an answer. Until recently, however, there have been few investigations into lungfish hearing. Writing in the *Journal of Experimental Biology*, Christensen *et al.*[1] report their findings about whether the African lungfish *Protopterus annectens* can detect sound, casting fresh light on our understanding of the hearing capabilities of the earliest tetrapods.

The earliest tetrapods seem not to have had a specialized apparatus that would enable terrestrial hearing[2], so to what extent could they pick up air-borne sound as they came onto land? Although there have been many studies of the hearing capacities of ray-finned fishes (actinopterygians), how relevant these findings are to early tetrapods has remained unclear, because ray-finned fishes are a separate branch of bony fishes (osteichthyans) from the lobe-finned fishes (sarcopterygians), the group to which tetrapods and lungfishes belong.

Lungfishes have no obvious adaptations for hearing — that is, they have no middle-ear cavity or a bone equivalent to the stapes bone in tetrapods, through which sound could be conveyed to the inner ear. However, they do have paired lungs, and Christensen and co-workers find that these are key to enabling the lungfish to detect sound.

Overcoming the obstacles to investigating sound detection by fishes is not simple. Ideally, the experiments should be done in open water to avoid the influence of the experimental set-up on the characteristics of the sound field. However, a sound field that can reasonably be used to examine the different aspects of fish hearing[3,4] can be established by creating a standing wave in a metal tube.

By placing lungfish at different locations within the tube, Christensen and colleagues

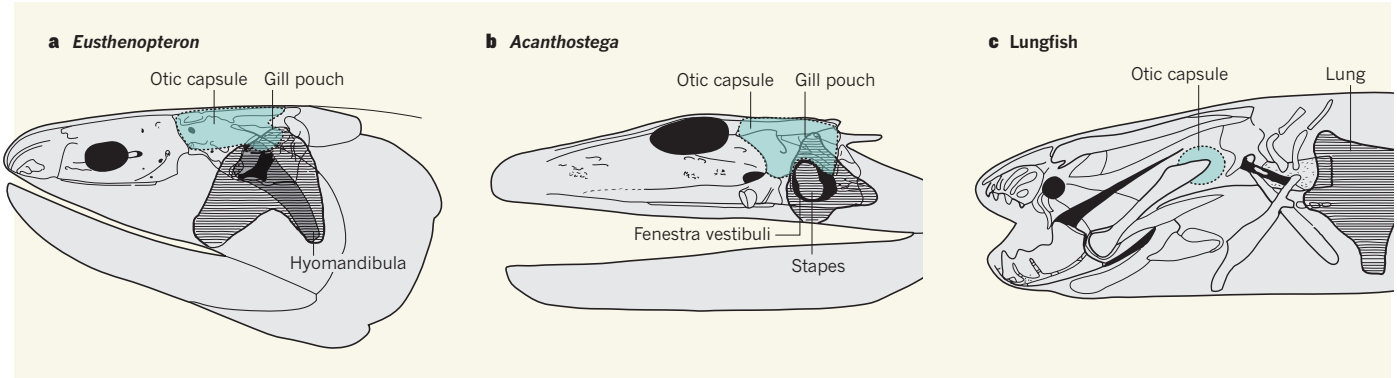


Figure 1 | Evolution of the hearing apparatus in tetrapods. **a**, In the extinct lobe-finned fish *Eusthenopteron*, a bone called the hyomandibula is associated with an air-filled gill pouch and articulates with the otic capsule (the ear region) to control movements of the lower jaw and other parts of the head and throat. **b**, *Acanthostega* is intermediate between lobe-finned fishes and the first tetrapods that were fully capable of coming onto land. A bone called the stapes (formed in part from a reduced-size hyomandibula) penetrates the otic capsule through an opening called the fenestra vestibuli. The stapes could transmit vibrations emanating from sound-induced pressure changes in the air-filled gill pouch to the otic capsule. **c**, Christensen *et al.*[1] report that the modern African lungfish, *Protopterus annectens*, uses its lung to transmit sound vibrations to the otic capsule, and provides a model for hearing in early tetrapods. (**a**, **b** adapted from ref. 10; **c** adapted from ref. 11.)

calibrated both pressure and particle motion in the set-up, to establish which of these components of the sound wave the fish were responding to. They show that the lungfish responds more strongly to the pressure generated by the sound than to particle motion. More specifically, it uses its air-filled lungs to convert pressure to particle motion in its lung that is then perceived by the inner ear. This is similar to the way in which ray-finned fishes use their swim bladder[5] (an internal gas-filled organ that allows a fish to control its buoyancy) for sound detection.

The researchers went on to show that lungfish can detect sound pressure waves propagated either through water or through the substrate (the material at the bottom of a lake or stream) and might even have a rudimentary capability to detect such waves in air, despite the absence of a direct anatomical connection between the lung and the inner ear. The groups' earlier work[6] had suggested that lungfish were unlikely to be able to detect pressure waves, but Christensen *et al.* obtained more positive results by using a modified version of the previously reported experimental set-up.

Their findings might have been predicted in the light of what is known about ray-finned fishes. But confirmation was necessary, and has major implications for the evolution of hearing in the earliest tetrapods. It suggests that, if lungfish are capable of sound detection without any obvious connection between an air bladder and the inner ear, then the presence of any such connection — even one not obviously adapted for hearing — would have made sound detection possible.

Both ray-finned and lobe-finned fishes are thought to have possessed air bladders early in their evolution, and may have used them in addition to gills for breathing. The swim bladder of ray-finned fishes is widely considered to have a common evolutionary origin with the lungs in lungfishes and tetrapods. All of the early bony fishes were also equipped with a bone called a hyomandibula that articulated with the ear capsule at a mobile joint and controlled ventilatory movements. It operated the pumping action of the gill chamber, throat and the buccal cavity (the mouth), drawing water or air into these spaces. Air could also pass into the air bladder by this mechanism. Lungfishes, however, lost the hyomandibula during their evolution, although they still breathe air using a similar, but elaborated, buccal-pumping mechanism.

It therefore seems that, even in the earliest osteichthyans, the proximity of the mobile hyomandibula to an air-filled chamber could have allowed pressure-induced vibrations to be transmitted to the inner ear. If air breathing was a primitive osteichthyan characteristic, these animals could, from the time of their origin, have detected sound propagated in water, through the substrate, or possibly even in air, and may have done it rather better than modern lungfishes.

Christensen and colleagues' discovery makes sense of what is known about the earliest tetrapods from the Late Devonian and Early Carboniferous periods (which together spanned from about 387 million to 323 million years ago). Two of the most obvious differences between the ear regions of early bony 'fish' and the descendent early 'tetrapods' are that, in the tetrapods, the hyomandibula had become modified into the stapes, which penetrated the braincase wall at an opening called the fenestra vestibuli (Fig. 1); and that the stapes had developed a structure called a stapedial footplate[7]. These changes must have marked at least some improvement in the transmission of sound waves to the inner ear, even though the stapes was not at that time a slender rod-like bone as it is in most land-dwelling tetrapods today. Rather, it was a bulky bone that was both relatively and absolutely much larger than in modern tetrapods[2] that have an eardrum and a middle-ear cavity, but was nonetheless capable of transmitting vibrations emanating from pressure changes in the air-filled gill pouch with which it was in contact.

The new discovery may also help to resolve an anomaly. One genus of Devonian tetrapod, *Ichthyostega*, had an ear region configured unlike that of any other known tetrapod. It seems to have had an air-filled chamber on each side of its head, roofed by thick walls formed by the skull, braincase and palate, but with a floor occupied in part by a thin, spoon-shaped stapes that articulated with the braincase and fenestra vestibuli[8]. This has been interpreted as an ear adapted for underwater hearing, yet other parts of *Ichthyostega*'s anatomy suggest that the animal had some adaptations for land locomotion[9]. We can now interpret the structure as an ear capable of hearing in both aquatic and terrestrial conditions. ■

Jennifer A. Clack *is at the University Museum of Zoology, University of Cambridge, Cambridge CB2 3EJ, UK.*
e-mail: j.a.clack@zoo.cam.ac.uk

1. Christensen, C. B., Christensen-Dalsgaard, J. & Madsen, P. T. *J. Exp. Biol.* **218,** 381–387 (2015).
2. Clack, J. A. *Brain Behav. Evol.* **50,** 198–212 (1997).
3. Hawkins, A. D. in *Perspectives on Auditory Research* (eds Popper, A. N. & Fay, R. R.) 247–267 (Springer, 2014).
4. Rogers, P. H., Hawkins, A. D., Popper, A. N., Fay, R. R. & Gray, M. D. in *The Effects of Noise on Aquatic Life, II* (eds Popper, A. N. & Hawkins, A. D.) (Springer, in the press).
5. Popper, A. N. & Fay, R. R. *Hear. Res.* **273,** 25–36 (2011).
6. Christensen-Dalsgaard, J., Brandt, C., Wilson, M., Wahlberg, M. & Madsen, P. T. *Biol. Lett.* **7,** 139–141 (2011).
7. Clack, J. A. *Nature* **369,** 392–394 (1994).
8. Clack, J. A. *et al. Nature* **425,** 65–69 (2003).
9. Pierce, S. E., Clack, J. A. & Hutchinson, J. R. *Nature* **486,** 523–526 (2012).
10. Clack, J. A. & Beneteau, A. *Gaining Ground: The Origin and Evolution of Tetrapods* 2nd edn (Indiana Univ. Press, 2012).
11. Long, J. A. *Mem. Assoc. Aust. Palaeontol.* **15,** 199–209 (1993).

doi:10.1038/nature14258

Defining the Anthropocene

Simon L. Lewis[1,2] & Mark A. Maslin[1]

Time is divided by geologists according to marked shifts in Earth's state. Recent global environmental changes suggest that Earth may have entered a new human–dominated geological epoch, the Anthropocene. Here we review the historical genesis of the idea and assess anthropogenic signatures in the geological record against the formal requirements for the recognition of a new epoch. The evidence suggests that of the various proposed dates two do appear to conform to the criteria to mark the beginning of the Anthropocene: 1610 and 1964. The formal establishment of an Anthropocene Epoch would mark a fundamental change in the relationship between humans and the Earth system.

Human activity has been a geologically recent, yet profound, influence on the global environment. The magnitude, variety and longevity of human-induced changes, including land surface transformation and changing the composition of the atmosphere, has led to the suggestion that we should refer to the present, not as within the Holocene Epoch (as it is currently formally referred to), but instead as within the Anthropocene Epoch[1–4] (Fig. 1). Academic and popular usage of the term has rapidly escalated[5,6] following two influential papers published just over a decade ago[1,2]. Three scientific journals focusing on the topic have launched: *The Anthropocene*, *The Anthropocene Review* and *Elementa*. The case for a new epoch appears reasonable: what matters when dividing geological-scale time is global-scale changes to Earth's status, driven by causes as varied as meteor strikes, the movement of continents and sustained volcanic eruptions. Human activity is now global and is the dominant cause of most contemporary environmental change. The impacts of human activity will probably be observable in the geological stratigraphic record for millions of years into the future[7], which suggests that a new epoch has begun[4].

Nevertheless, some question the types of evidence[8,9], because to define a geological time unit, formal criteria must be met[10,11]. Global-scale changes must be recorded in geological stratigraphic material, such as rock, glacier ice or marine sediments (see Box 1). At present, there is no formal agreement on when the Anthropocene began, with proposed dates ranging from before the end of the last glaciation to the 1960s. Such different meanings may lead to misunderstandings and confusion across several disciplines. Furthermore, unlike other geological time unit designations, definitions will probably have effects beyond geology. For example, defining an early start date may, in political terms, 'normalize' global environmental change. Meanwhile, agreeing a later start date related to the Industrial Revolution may, for example, be used to assign historical responsibility for carbon dioxide emissions to particular countries or regions during the industrial era. More broadly, the formal definition of the Anthropocene makes scientists arbiters, to an extent, of the human–environment relationship, itself an act with consequences beyond geology. Hence, there is more interest in the Anthropocene than other epoch definitions. Nevertheless, evidence will define whether the geological community formally ratifies a human-activity-induced geological time unit.

We therefore review human geology in four parts. First, we summarize the geologically important human-induced environmental impacts. Second, we review the history of naming the epoch that modern human societies live within, to provide insights into contemporary Anthropocene-related debates. Third, we assess environmental changes caused by human activity that may have left global geological markers consistent with the formal criteria that define geological epochs. Fourth, we highlight the

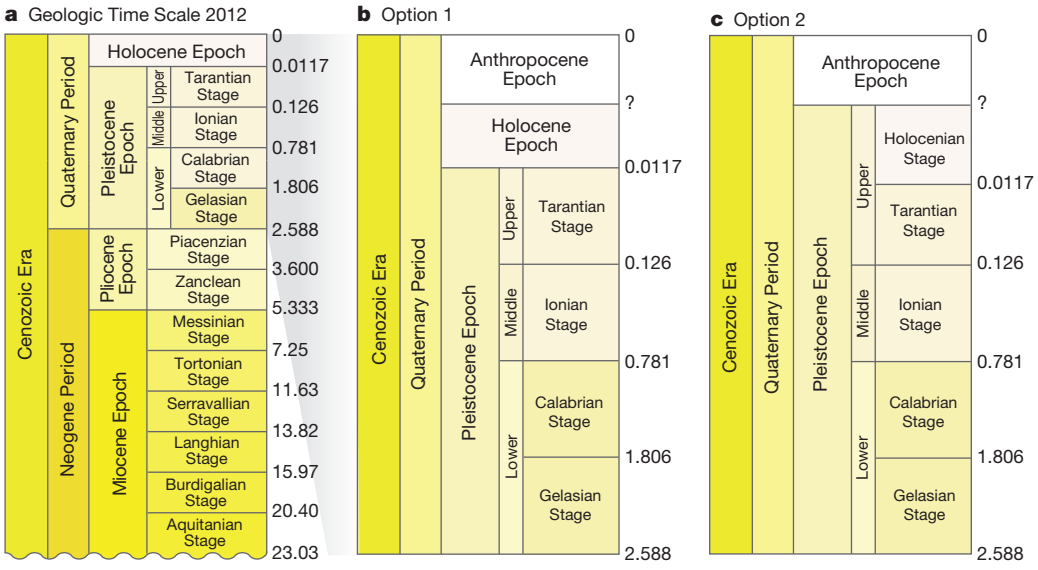


Figure 1 | Comparison of the current Geologic Time Scale[10] (GTS2012), with two alternatives. **a**, GTS2012, with boundaries marked in millions of years (ref. 10). **b**, **c**, The alternatives include a defined Anthropocene Epoch following either the Holocene (**b**) or directly following the Pleistocene (**c**). Defining the Anthropocene as an epoch requires a decision as to whether the Holocene is as distinct as the Anthropocene and Pleistocene; retaining it or not distinguishes between **b** and **c**. The question mark represents the current debate over the start of the Anthropocene, assuming it is formally accepted as an epoch (see Box 1, Fig. 2). Colour coding is used according to the Commission for the Geological Map of the World[10], except for the Anthropocene.

[1]Department of Geography, University College London, Gower Street, London, WC1E 6BT, UK. [2]School of Geography, University of Leeds, Leeds, LS2 9JT, UK.

BOX 1

Dividing geological time

Geological time is divided into a hierarchical series of ever-finer units (Fig. 1a). The present, according to *The Geologic Time Scale 2012*[10], is in the Holocene Epoch (Greek for 'entirely recent'; started 11,650 yr BP), within the Quaternary Period (started 2.588 million years ago), within the Cenozoic Era ('recent life'; started 66 million years ago) of the Phanerozoic Eon ('revealed life'; started 541 million years ago). Divisions represent differences in the functioning of Earth as a system and the concomitant changes in the resident life-forms. Larger differences result in classifications at higher unit-levels.

Formally, geological time units are defined by their lower boundary, that is, their beginning. Boundaries are demarcated using a GSSP, or if good candidate GSSPs do not exist, by an agreed date, termed a GSSA[10]. For a GSSP, a 'stratotype section' refers to a portion of material that develops over time (rock, sediment, glacier ice), and 'point' refers to the location of the marker within the stratotype. Each 'golden spike' is a single physical manifestation of a change recorded in a stratigraphic section, often reflecting a global-change phenomenon. GSSP markers are then complemented by a series of correlated changes, also recorded stratigraphically, termed auxiliary stratotypes, indicating widespread changes to the Earth system occurring at that time[10]. An exemplary GSSP is the Cretaceous–Paleogene period-level boundary, and the start of the Cenozoic Era, when non-avian dinosaurs declined to extinction and mammals radically increased in variety and abundance. The GSSP boundary marker is the peak in iridium—a residual of bolide impact with Earth—in rock dated at 66 million years ago, located at El Kef, Tunisia[10].

The widespread appearance of new species can also be used as GSSP boundary markers; for example, the Ordovician–Silurian period-level boundary, 443.8 million years ago, is marked by the appearance of a distinct planktonic graptolite, *Akidograptus ascensus* (a now-extinct hemichordate)[10]. From an Anthropocene perspective this example shows that the GSSP primary marker chosen as a boundary indicator may be of limited importance compared to the other events taking place that collectively show major changes to Earth at that time[67].

Formally, a GSSP must have (1) a principal correlation event (the marker), (2) other secondary markers (auxiliary stratotypes), (3) demonstrated regional and global correlation, (4) complete continuous sedimentation with adequate thickness above and below the marker, (5) an exact location—latitude, longitude and height/depth—because a GSSP can be located at only one place on Earth, (6) be accessible, and (7) have provisions for GSSP conservation and protection[10].

Alternatively, following a survey of the stratigraphic evidence, a GSSA date may be agreed by committee to mark a time unit boundary. GSSAs are typical in the Precambrian (>541 million years ago) because well-defined geological markers and clear events are less obvious further back in time[10]. Regardless of the marker type, formally ratifying a new Anthropocene Epoch into the GTS would first require a positive recommendation from the Anthropocene Working Group of the Subcommission of Quaternary Stratigraphy, followed by a supermajority vote of the International Commission on Stratigraphy, and finally ratification by the International Union of Geological Sciences[10] (see ref. 11 for full details).

advantages and disadvantages of the few global markers that may indicate a date to define the beginning of the Anthropocene. By consolidating research from disparate fields and the emerging Anthropocene-specific literature we aim to constrain the number of possible Anthropocene start dates, highlight areas requiring further research, and assist in moving towards an evidence-based decision on the possible ratification of a new Anthropocene Epoch.

The geological importance of human actions

Human activity profoundly affects the environment, from Earth's major biogeochemical cycles to the evolution of life. For example, the early-twentieth-century invention of the Haber–Bosch process, which allows the conversion of atmospheric nitrogen to ammonia for use as fertilizer, has altered the global nitrogen cycle so fundamentally that the nearest suggested geological comparison refers to events about 2.5 billion years ago[12]. Human actions have released 555 petagrams of carbon (where 1 Pg = 10^{15} g = 1 billion metric tons) to the atmosphere since 1750, increasing atmospheric CO_2 to a level not seen for at least 800,000 years, and possibly several million years[13,14], thereby delaying Earth's next glaciation event[15]. The released carbon has increased ocean water acidity at a rate probably not exceeded in the last 300 million years[16].

Human action also affects non-human life. Global net primary productivity appears to be relatively constant[17]; however, the appropriation of 25–38% of net primary productivity for human use[17,18] reduces the amount available for millions of other species on Earth. This land-use conversion to produce food, fuel, fibre and fodder, combined with targeted hunting and harvesting, has resulted in species extinctions some 100 to 1,000 times higher than background rates[19], and probably constitutes the beginning of the sixth mass extinction in Earth's history[19]. Species removals are non-random, with greater losses of large-bodied species from both the land and the oceans. Organisms have been transported around the world, including crops, domesticated animals and pathogens on land. Similarly, boats have transferred organisms among once-disconnected oceans. Such movement has led to a small number of extraordinarily common species, new hybrid species[20], and a global homogenization of Earth's biota. Ostensibly, this change is unique since Pangaea separated about 200 million years ago[21], but such trans-oceanic exchanges probably have no geological analogue.

Furthermore, human actions may well constitute Earth's most important evolutionary pressure[22,23]. The development of diverse products, including antibiotics[22], pesticides[22,24], and novel genetically engineered organisms[24], alongside the movement of species to new habitats[25], intense harvesting[23] and the selective pressure of higher air temperatures resulting from greenhouse gas emissions, are all likely to alter evolutionary outcomes[22–25]. Considered collectively, there is no geological analogue[22]. Furthermore, given that the lifespan of a species is typically 1–10 million years, the rates of anthropogenic environmental change in the near future may exceed the rates of change encountered by many species in their evolutionary history. Human activity has clearly altered the land surface, oceans and atmosphere, and re-ordered life on Earth.

Historical human geology

Human-related geological time units have a long history[26]. In 1778 Buffon published an early attempt to describe Earth's history, allocating a human epoch to be Earth's seventh and final epoch, paralleling the seven-day creation story[27]. By the nineteenth century, divine intervention was receding from consideration as a geological force. In 1854 the Welsh geologist and professor of theology, Thomas Jenkyn, appears to have first published the idea of an explicitly evidence-based human geological time unit in a series of widely disseminated geology lessons[28–30]. He describes the then present day as "the human epoch" based on the likely future fossil record[28]. In his final lecture he wrote, "All the recent rocks, called in our last lesson Post-Pleistocene, might have been called Anthropozoic, that is, human-life rocks."[29]. Similarly, the Reverend Haughton's 1865 Manual of Geology describes the Anthropozoic as the "epoch in which we live"[31], as did the Italian priest and geologist Antonio Stoppani a decade later[32]. Meanwhile in the USA, the geology professor James Dwight Dana's then-popular 1863 Manual of Geology[33] extensively refers to the "Age of Mind and Era of Man" as the youngest geological time, as did many of his US contemporaries[34].

In 1830 Charles Lyell had proposed that contemporary time be termed the Recent epoch[35] on the basis of three considerations: the end of the last glaciation, the then-believed coincident emergence of humans, and the

rise of civilizations[26,35]. In the 1860s, the French geologist Paul Gervais made Lyell's term international, coining the term Holocene, derived from the Greek for 'entirely recent'. Thus, most nineteenth-century geological textbooks feature humans as part of the definition of the most recent geological time units. Critically, there was little discussion about any of these terms—Recent, Holocene or Anthropozoic—probably because each represented the same conceptual model and broad agreement that humans were part of the definition of the contemporary geological epoch. However, the wider written records of these often deeply religious men show that a separate human epoch was likely to have been more strongly influenced by theological concerns—in particular, separating *Homo sapiens* from other animals and retaining humans at the apex of life on Earth—than by the appraisal of stratigraphic evidence.

In the twentieth century, geologists in the West increasingly used the term Holocene for the current epoch, and Quaternary for the period. Meanwhile, in 1922 the Russian geologist Aleksei Pavlov described the present day as part of an "Anthropogenic system (period) or Anthropocene"[36]. The Ukrainian geochemist Vladimir Vernadsky then brought to widespread attention the idea that the biosphere, combined with human cognition, had created the Noösphere (from the Greek for mind), with humans becoming a geological force[37]. The term Noösphere was not well used, but non-Western scientists often used anthropogenic geological time units. The Russian term was anglicized as both Anthropogene and Anthropocene[36], sometimes creating confusion. The East–West differences in usage may have been due to differing political ideologies: an orthodox Marxist view of the inevitability of global collective human agency transforming the world politically and economically requires only a modest conceptual leap to collective human agency as a driver of environmental transformation. Again there was little broad interest in the various terms. The Holocene became the official term within the Geologic Time Scale (GTS; Fig. 1)[10,38], with its implication that the current interglacial differs from the previous Pleistocene interglacials owing to the influence of humans. It has therefore been argued that an Anthropocene Epoch is not required, given that some human influence is already contained within the definition of the Holocene Epoch[9]. Alternatively, defining the Anthropocene would deprive the Holocene Epoch of its ostensibly unique feature—humans—suggesting that the Holocene as an epoch may not be required.

The views of nineteenth- and twentieth-century scientists illustrate the influence of the dominant contemporary concerns on geological debates. Today's scientists may also not be immune to such influences. For example, a key concern for scientists and others is the central role of technology in modern society and its environmental impacts. Crutzen and Stoermer[1] originally proposed that the start of the Anthropocene should be coincident with the beginning of the Industrial Revolution and James Watt's 1784 refinement of the steam engine. Others followed, including stratigraphers, suggesting that 1800 should be the beginning of the Anthropocene[39,40], despite a lack of corresponding global geological markers, and the presence of well-known stratigraphic evidence suggestive of different dates, such as the radionuclide fallout from mid-twentieth-century nuclear weapons tests. Care is needed to ensure that the dominant culture of today's scientists does not subconsciously influence the assessment of stratigraphic evidence.

A human golden spike

Defining the beginning of the Anthropocene as a formal geologic unit of time requires the location of a global marker of an event in stratigraphic material, such as rock, sediment, or glacier ice, known as a Global Stratotype Section and Point (GSSP), plus other auxiliary stratigraphic markers indicating changes to the Earth system. Alternatively, after a survey of the stratigraphic evidence, a date can be agreed by committee, known as a Global Standard Stratigraphic Age (GSSA). GSSPs, known as 'golden spikes', are the preferred boundary markers[10] (see Box 1).

Generally, geologists have used temporally distant changes in multiple stratigraphic records to delimit major changes in the Earth system and thereby geological time units, for example, the appearance of new species as fossils within rocks, coupled with other temporally coincident changes. Perhaps the most useful GSSP example when considering a possible Anthropocene GSSP is that marking the beginning of the most recent epoch, the Holocene[38], because some similar choices and difficulties were faced. These include: not relying on solid aggregate mineral deposits ('rock') for the boundary; an event horizon largely lacking fossils (although fossils are used to recognize Holocene deposits); the need for very precise GSSP dating of events in the recent past; and how to formalize a time unit that extends to the present and thereby implicitly includes a view of the future.

Depending on the parameter considered, the current interglacial took decades to millennia to unfold, as global climate, atmospheric chemistry and the distribution of plant and animal species all altered. From these changes a single dated level within a single stratigraphic record was required to be chosen as a GSSP primary marker (Box 1; Fig. 2). Thus, formally, the Holocene is marked by an abrupt shift in deuterium (^{2}H) excess values at a depth of 1,492.25 m in the NorthGRIP Greenland ice core, dated 11,650 ± 99 yr BP (before present, where 'present' is defined to be 1950)[38]. This corresponds to the first signs of predominantly Northern Hemisphere climatic warming at the end of the Younger Dryas/ Greenland Stadial 1 cold period[38] (Fig. 2). Five further auxiliary stratotypes (four lakes and one marine sediment) showing clear correlated changes across the boundary complement the GSSP, consistent with the occurrence of global changes to the Earth system[38]. The requirements for a formal definition of the start of the Anthropocene are similar: a clear, datable marker documenting a global change that is recognizable in the stratigraphic record, coupled with auxiliary stratotypes documenting long-term changes to the Earth system.

Defining the Anthropocene presents a further challenge. Changes to the Earth system are not instantaneous. However, even spatially heterogeneous and diachronous (producing similar stratigraphic material varying in age) changes appear near-instantaneous when viewed millions of years after the event, especially as time-lags often fall within the error range of the dating techniques. In contrast, Anthropocene deposits are commonly dated on decadal or annual scales, so that all changes will appear diachronous, to some extent, from today's perspective (but not from far in the future)[11,41]. Judgement will be required to assess whether the time-lags following events and their significant global impacts are too long to be of use when defining any Anthropocene GSSP.

Several approaches have been put forward to define when the Anthropocene began, including those focusing on the impact of fire[42], pre-industrial farming[43–45], sociometabolism[46], and industrial technologies[1,39,40,41,47], but the relative merits of the evidence for various starting dates have not been systematically assessed against the requirements of a golden spike. Below, we review the major events in human history and pre-history and their impact on stratigraphic records. We focus on continuous stratigraphic material that may yield markers consistent with a GSSP (lake and marine sediments, glacier ice) and on the types of chemical, climatic and biological changes used to denote other epoch boundaries further in the past. We proceed chronologically forward in time, presenting the reason why each event was originally proposed, evaluate the existence of stratigraphic markers, and assess whether the event provides a potential GSSP. The hypotheses and evidence are summarized in Table 1. Following the evidence review we briefly consider the relative merits of the differing events that probably fulfil the GSSP criteria, and assess related GSSA dates.

Pleistocene human impacts

The first major impacts of early humans on their environment was probably the use of fire. Fossil charcoal captures these events from the Early Pleistocene Epoch[42,48]. However, fires are inherently local events, so they do not provide a global GSSP. The next suggested candidate is the Megafauna Extinction between 50,000 and 10,000 years ago, given that other epoch boundaries have been defined on the basis of extinctions or on the resultant newly emerging species[10]. Overall, during the Megafauna Extinction about half of all large-bodied mammals worldwide, equivalent to 4% of all mammal species, were lost[49]. The losses were not evenly distributed: Africa lost 18%, Eurasia lost 36%, North

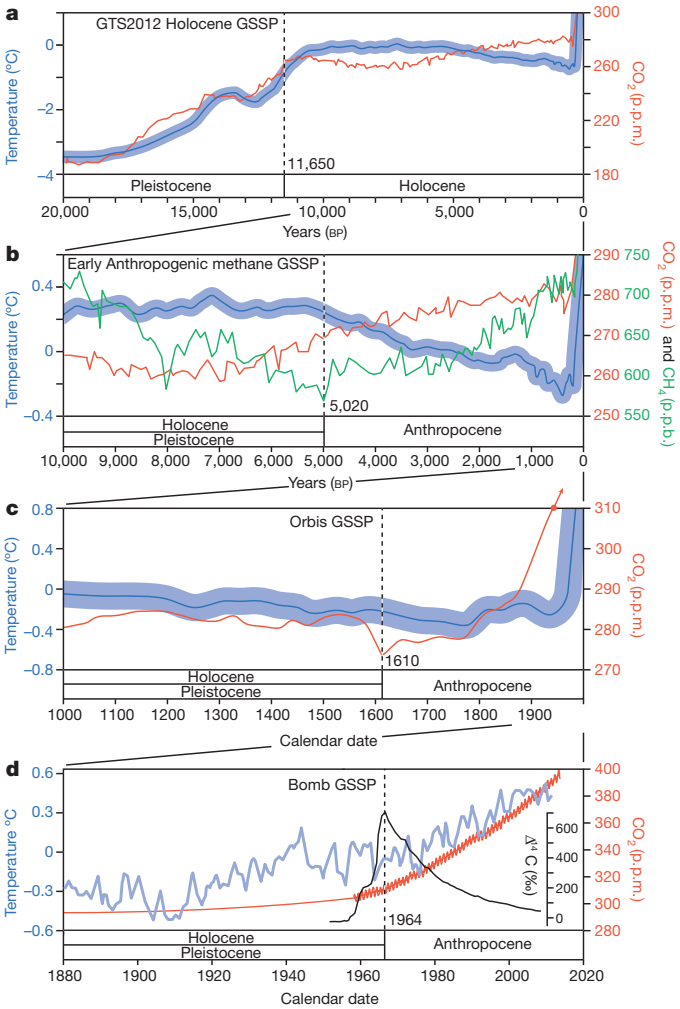


Figure 2 | Defining the beginning of the Anthropocene. **a**, Current GTS2012 GSSP boundary between the Pleistocene and Holocene[38] (dashed line), with global temperature anomalies (relative to the early Holocene average over the period 11,500 BP to 6,500 BP)[112] (blue), and atmospheric carbon dioxide composite[113] on the AICC2012 timescale[114] (red). **b**, Early Anthropogenic Hypothesis GSSP suggested boundary (dashed line), which posits that early extensive farming impacts caused global environmental changes, defined here by the inflection and lowest level of atmospheric methane (in parts per billion, p.p.b.) from the GRIP ice core[59] (green), with global temperature anomalies (relative to the average over the period 1961 to 1990)[115] (blue), and atmospheric carbon dioxide[113] (red). **c**, Orbis GSSP suggested boundary (dashed line), representing the collision of the Old and New World peoples and homogenization of once distinct biotas, and defined by the pronounced dip in atmospheric carbon dioxide (dashed line) from the Law Dome ice core[75,76] (blue), with global temperature data anomalies (relative to the average over the period 1961 to 1990)[115] (red). **d**, Bomb GSSP suggested boundary (dashed line), characterized by the peak in atmospheric radiocarbon from annual tree-rings (black)[103] (the $\Delta^{14}C$ value is the relative difference between the absolute international standard (base year 1950) and sample activity corrected for the time of collection and $\delta^{13}C$), with atmospheric carbon dioxide from Mauna Loa, Hawaii, post-1958[116], and ice core records pre-1958[75,76] (red), and global temperature anomalies (relative to the average over the period 1961 to 1990)[116] (blue).

America lost 72%, South America lost 83%, and Australia lost 88% of their large-bodied mammalian genera[50,51]. So the Megafauna Extinction was actually a series of events on differing continents at differing times and therefore lacks the required precision for an Anthropocene GSSP marker.

Origins and impacts of farming

The development of agriculture causes long-lasting anthropogenic environmental impacts as it replaces natural vegetation, and thereby increases species extinction rates, and alters biogeochemical cycles. Agriculture had multiple independent origins: first occurring about 11,000 years ago in southwest Asia, South America and north China; between 6,000–7,000 years ago in Yangtze China and Central America; and 4,000–5,000 years ago in the savanna regions of Africa, India, southeast Asia, and North America[52]. Thus, the increasing presence of fossil pollen from domesticated plants in sediment is too local and lacking in global synchrony to form a GSSP marker. Critically, for the Holocene GSSP, auxiliary markers within stratigraphic material did not include any human-derived markers[38], illustrating the lack of anthropogenic impacts at that time. Long-lasting cultural evidence related to agriculture is similarly constrained. Although ceramics are datable and preserved in stratigraphic records (for example, the mineral mullite[41]), they appeared in Africa before agriculture, while early southwest Asian farming cultures did not produce ceramics. Similarly, anthropogenically formed soils, derived from intensive farmland management, have also been suggested as a marker of the Anthropocene[53]. Although these soils are widespread, like vegetation clearance, they are highly diachronous over about 2,000 yr, thus excluding their use as a GSSP marker[54].

A series of Neolithic revolutions resulted in the majority of *Homo sapiens* becoming agriculturalists to some extent by around 8,000 yr BP, rising to a maximum of about 99% by about 500 yr BP[46]. The Early Anthropogenic Hypothesis posits that the current interglacial was similar to the previous seven interglacial periods until around 8,000 yr BP[43,55]. By comparison with the closest astronomical analogue of the current interglacial (795,000–780,000 yr BP)[55], atmospheric CO_2 should have continued to decline after 8,000 yr BP, eventually reaching about 240 parts per million (p.p.m.), and the onset of glaciation should have begun[43,55]. However, by 6,000–8,000 yr BP, farmers' conversion of high-carbon storage vegetation (forest, woodland, woody savanna) to crops and grazing lands, plus associated fire impacts, may have increased atmospheric CO_2 levels, and postponed this new glaciation[43] (Fig. 2). Thus, the lowest level of CO_2 within an ice core record could, in principle, provide a golden spike, but the CO_2 record lacks a distinct inflection point at this time (Fig. 2). Furthermore, the evidence that human activity was responsible for the gradual increase in CO_2 after 6,000 yr BP is extensively debated[43,56–58].

Methane provides a clearer inflection point, which may provide a possible GSSP at 5,020 yr BP, the date of the lowest methane value recorded in the GRIP ice core[59] (Fig. 2). Archaeological evidence suggests that the inflection is caused by rice cultivation in Asia and the expansion of populations of domesticated ruminants. Comparisons of changes in atmospheric methane from the current and past interglacials[43], and some methane $\delta^{13}C$ value evidence[60], also suggest a human cause. However, a model study suggests that orbital forcing altering methane emissions from tropical wetlands may be responsible[61]. Auxiliary markers could include stone axes and fossilized domesticated crop pollen and ruminant remains, but these do not provide temporally well-correlated markers that collectively document globally synchronous changes to the Earth system.

Collision of the Old and New Worlds

The arrival of Europeans in the Caribbean in 1492, and subsequent annexing of the Americas, led to the largest human population replacement in the past 13,000 years[62], the first global trade networks linking Europe, China, Africa and the Americas[63,64], and the resultant mixing of previously separate biotas, known as the Colombian Exchange[63,64]. One biological result of the exchange was the globalization of human foodstuffs. The New World crops maize/corn, potatoes and the tropical staple manioc/cassava were subsequently grown across Europe, Asia and Africa. Meanwhile, Old World crops such as sugarcane and wheat were planted in the New World. The cross-continental movement of dozens of other food species (such as the common bean, to the New World), domesticated animals (such as the horse, cow, goat and pig, all to the Americas) and human commensals (the black rat, to the Americas), plus accidental transfers (many species of earth worms, to North America; American mink to Europe) contributed to a swift, ongoing, radical reorganization of life on Earth without geological precedent.

Table 1 | Potential start dates for a formal Anthropocene Epoch

Event	Date	Geographical extent	Primary stratigraphic marker	Potential GSSP date*	Potential auxiliary stratotypes
Megafauna extinction	50,000–10,000 yr BP	Near-global	Fossil megafauna	None, diachronous over ~40,000 yr	Charcoal in lacustrine deposits
Origin of farming	~11,000 yr BP	Southwest Asia, becoming global	Fossil pollen or phytoliths	None, diachronous over ~5,000 yr	Fossil crop pollen, phytoliths, charcoal
Extensive farming	~8,000 yr BP to present	Eurasian event, global impact	CO_2 inflection in glacier ice	None, inflection too diffuse	Fossil crop pollen, phytoliths, charcoal, ceramic minerals
Rice production	6,500 yr BP to present	Southeast Asian event, global impact	CH_4 inflection in glacier ice	5,020 yr BP CH_4 minima	Stone axes, fossil domesticated ruminant remains
Anthropogenic soils	~3,000–500 yr BP	Local event, local impact, but widespread	Dark high organic matter soil	None, diachronous, not well preserved	Fossil crop pollen
New–Old World collision	1492–1800	Eurasian–Americas event, global impact	Low point of CO_2 in glacier ice	1610 CO_2 minima	Fossil pollen, phytoliths, charcoal, CH_4, speleothem $\delta^{18}O$, tephra†
Industrial Revolution	1760 to present	Northwest Europe event, local impact, becoming global	Fly ash from coal burning	~1900 (ref. 94); diachronous over ~200 yr	^{14}N:^{15}N ratio and diatom composition in lake sediments
Nuclear weapon detonation	1945 to present	Local events, global impact	Radionuclides (^{14}C) in tree-rings	1964 ^{14}C peak§	^{240}Pu: ^{239}Pu ratio, compounds from cement, plastic, lead and other metals
Persistent industrial chemicals	~1950 to present	Local events, global impact	For example, SF_6 peak in glacier ice	Peaks often very recent so difficult to accurately date§	Compounds from cement, plastic, lead and other metals

For compliance with a Global Stratotype Section and Point (GSSP) definition, a clearly dated global marker is required, backed by correlated auxiliary markers that collectively indicate global and other widespread and long-term changes to the Earth system. BP, before present, where present is defined as calendar date 1950.
* Requires a specific date for a GSSP primary marker. †From Huaynaputina eruption in 1600 (refs 78, 79).
§ Peak, rather than earliest date of detection selected, because earliest dates reflect available detection technology, are more likely influenced by natural background geochemical levels[101], and will be more affected by the future decay of the signal, than peak values.

In terms of stratigraphy, the appearance of New World plant species in Old World sediments—and vice versa—may provide a common marker of the Anthropocene across many deposits because pollen is often well preserved in marine and lake sediments. For example, pollen of New World native *Zea mays* (maize/corn), which preserves very well[41], first appears in a European marine sediment core in 1600[65]. The European Pollen Database lists a further 70 lake and marine sediment cores containing *Zea mays* after this date. Phytoliths can similarly record such range expansions[66]. Specifically, the transcontinental range extension of at least one Old World species into the New World (banana, as phytoliths in Central and tropical South America sediments) and a second species from the New World expanding into the Old World (maize/corn, as pollen preserved in sediments in Eurasia and Africa) together constitute a unique signature in the stratigraphic record. This transcontinental range expansion—stratigraphically marking before and after an event—is comparable to the use of the appearance of new species as boundary markers in other epoch transitions[49,67].

Besides permanently and dramatically altering the diet of almost all of humanity, the arrival of Europeans in the Americas also led to a large decline in human numbers. Regional population estimates sum to a total of 54 million people in the Americas in 1492[68], with recent population modelling estimates of 61 million people[58]. Numbers rapidly declined to a minimum of about 6 million people by 1650 via exposure to diseases carried by Europeans, plus war, enslavement and famine[58,63,68,69]. The accompanying near-cessation of farming and reduction in fire use resulted in the regeneration of over 50 million hectares of forest, woody savanna and grassland with a carbon uptake by vegetation and soils estimated at 5–40 Pg within around 100 years[58,70–72]. The approximate magnitude and timing of carbon sequestration suggest that this event significantly contributed to the observed decline in atmospheric CO_2 of 7–10 p.p.m. (1 p.p.m. CO_2 = 2.1 Pg of carbon) between 1570 and 1620 documented in two high-resolution Antarctic ice core records[73–76] (Fig. 2 and Box 2). This dip in atmospheric CO_2 is the most prominent feature, in terms of both rate of change and magnitude, in pre-industrial atmospheric CO_2 records over the past 2,000 years[75] (Fig. 2).

On the basis of the movement of species, atmospheric CO_2 decline and the resulting climate-related changes within various stratigraphic records, we propose that the 7–10 p.p.m. dip in atmospheric CO_2 to a low point of 271.8 p.p.m. at 285.2 m depth of the Law Dome ice core[75], dated 1610 (±15 yr; refs 75, 76), is an appropriate GSSP marker (Fig. 2). Auxiliary stratotypes could include: the first occurrence of a cross-ocean range extension in the fossil record (*Zea mays*, in 1600[65]) plus a range of deposits showing distinct changes at that time, including tephra[77,78] and other signatures from the 1600 Huaynaputina eruption detected at both poles and in the tropics[77–79]; charcoal reductions in deposits in the Americas[71] and globally[80]; decreases in atmospheric methane, enrichment of methane $\delta^{13}C$, and decreases in carbon monoxide in Antarctic ice cores[60,81–84]; pollen in lacustrine sediments showing vegetation regeneration[85]; proxies indicating anomalous Arctic sea-ice extent[86]; changing $\delta^{18}O$ derived from speleothems from caves in China and Peru[14] and other studies noting changes coincident with 1600 and the coolest part of the Little Ice Age (1594–1677; ref. 87), a relatively synchronous global event noted in geologic deposits worldwide[87].

The impacts of the meeting of Old and New World human populations—including the geologically unprecedented homogenization of Earth's biota[63,64]—may serve to mark the beginning of the Anthropocene. Although it represents a major event in world history[62–64,88], the collision of the Old and New Worlds has not been proposed previously, to our knowledge, as a possible GSSP. We suggest naming the dip in atmospheric CO_2 the 'Orbis spike' and the suite of changes marking 1610 as the beginning of the Anthropocene the 'Orbis hypothesis', from the Latin for world, because post-1492 humans on the two hemispheres were connected, trade became global, and some prominent social scientists refer to this time as the beginning of the modern 'world-system'[89].

Industrialization

The beginning of the Industrial Revolution has often been suggested as the beginning of the Anthropocene, because accelerating fossil fuel use and coupled rapid societal changes herald something important and unique in human history[1–4,39]. Yet humans have long been engaging in industrial-type production, such as metal utilization from around 8,000 yr BP onwards, with attendant pollution[90]. Elevated mercury records are documented at around 3,400 yr BP in the Peruvian Andes[91], while the impacts of Roman Empire copper smelting are detectable in a Greenland ice core at around 2,000 yr BP[92]. This metal pollution, like other examples predating the Industrial Revolution, is too local and diachronous to provide a golden spike.

BOX 2

Origins of the 1610 decrease in atmospheric CO_2

Is the CO_2 decline real?

Two independent high-resolution Antarctic ice core records from the Law Dome and the Western Antarctic Ice Sheet show a reduction in atmospheric CO_2 of 7–10 p.p.m. between 1570 and 1620[73–75] (Fig. 2). A smaller CO_2 decrease is also observed in less highly resolved Antarctic cores[117,118]. The decline exceeds the measurement error of the cores, 1–2 p.p.m., and experiments suggest that it does not result from *in situ* changes within the ice core[119].

Did human activity cause the decline?

The arrival of Europeans in the Americas led to a catastrophic decline in human numbers, with about 50 million deaths between 1492 and 1650, according to several independent sources[58,63,68,69]. Contemporary field observations of soil[120] and vegetation[121] carbon dynamics following agriculture abandonment suggest that about 65 million hectares (that is, 50 million people × 1.3 hectares per person) would sequester 7–14 Pg of carbon over 100 years (that is, 100–200 Mg of carbon per hectare total uptake, above- and below-ground). Reduction in fire use for land management would additionally increase carbon uptake outside farmed areas. Studies using a variety of methods report broadly consistent estimates[58,70–72] of carbon uptake by vegetation of 5–40 Pg (2.1 Pg of carbon = 1 p.p.m. atmospheric CO_2 over shorter timescales, lessening over time[127]). Given that maximum human mortality rates were not reached for some decades after 1492[62,63], and maximum carbon uptake would take place 20–50 yr after farming abandonment, peak carbon sequestration would occur approximately between 1550 and 1650.

Some model studies spanning thousands of years find a net land surface carbon uptake spanning 1500–1650 across the Americas[58], while others do not[122]. However, in general, evidence from such studies weakly constrain the problem because Holocene carbon cycle modelling is designed to investigate changes associated with long-acting slow processes (carbon uptake by peat or coral reefs) and feedback mechanisms (oceanic outgassing, oceanic uptake and CO_2 fertilization of vegetation), and probably poorly represent the short period of the CO_2 dip (for example, ref. 57). For example, a study calculating a net zero impact of the cessation of farming in the Americas[122] included a large soil carbon flux to the atmosphere, which contradicts field evidence[120,123], and had the effect of offsetting the uptake from growing trees[122]. Carbon cycle models with robust representations of land-use change and subsequent vegetation regeneration following the Americas population catastrophe will be required to improve estimates of carbon uptake compared with carbon accounting studies.

The approximate magnitude and timing of carbon sequestration make the population decline in the Americas the most likely cause of the observed decline in atmospheric CO_2. Atmospheric[74,124,125] and tropical marine $\delta^{13}C$ analyses[126] also support uptake of CO_2 by vegetation rather than oceanic uptake. The 1600 Huaynaputina eruption in Peru[78,79] probably exacerbated the CO_2 minima, and a lagged oceanic outgassing in response to the land carbon uptake probably contributed to the fast rebound of atmospheric CO_2 after 1610[127]. In addition, multi-proxy reconstructions of temperature indicate that, after accounting for both solar and volcanic radiative forcing, additional terrestrial carbon uptake is required to explain temperature declines over the 1550–1650 period[107]. This is consistent with uptake by vegetation following the population crash in the Americas[107].

Definitions of the Industrial Revolution give an onset date anywhere between 1760 and 1880, beginning as an event local to northwest Europe[88]. Given the initial slow spread of coal use, ice core records show little impact on global atmospheric CO_2 concentration until the nineteenth century, and then they show a relatively smooth increase rather than an abrupt change, precluding this as a GSSP marker (Fig. 2). Similarly, other associated changes, including methane and nitrate[15], products of fossil fuel burning (including spherical carbonaceous particles[93] and magnetic fly ash[94]) plus resultant changes in lake sediments[95,96] alter slowly as the use of fossil fuels increased over many decades. Lead, which was once routinely added to vehicle fuels, has been proposed as a possible marker, because leaded fuel was almost globally used and is now banned[97]. However, peak lead isotope ratio values from this source in sediments and other deposits vary from 1940 to after 1980, limiting the utility of this marker. The Industrial Revolution thus provides a number of markers spreading from northwest Europe to North America and expanding worldwide since about 1800, although none provides a clear global GSSP primary marker.

The Great Acceleration

Since the 1950s the influence of human activity on the Earth system has increased markedly. This 'Great Acceleration' is marked by a major expansion in human population, large changes in natural processes[3,12,98], and the development of novel materials from minerals to plastics to persistent organic pollutants and inorganic compounds[41,47,97]. Among these many changes the global fallout from nuclear bomb tests has been proposed as a global event horizon marker[41,47]. The first detonation was in 1945, with a peak in atmospheric testing from the late 1950s to early 1960s, followed by a rapid decline following the Partial Test Ban Treaty in 1963 and later agreements, such that only low test levels continue to the present day (Fig. 2). A resulting distinct peak in radioactivity is recorded in high-resolution ice cores, lake and salt marsh sediments, corals, speleothems and tree-rings from the early 1950s onwards, declining in the late 1960s[15,99]. The clearest signal is from atmospheric ^{14}C, seen in direct air measurements and captured by tree-rings and glacier ice, which reaches a maximum in the mid- to high-latitude Northern Hemisphere at 1963–64 and a year later in the tropics[100]. Although ^{14}C has a relatively short half-life (5,730 years), elevated levels will persist long enough to be useable for several generations of geologists in the future.

While recognizing that many apparently novel industrially produced chemicals are occasionally produced in small quantities naturally[101], chemical signatures from long-lived well-mixed gases in glacier ice or sediments may also meet GSSP criteria. Potential long-lived gases are the halogenated gases, such as SF_6, C_2F_6, CF_4 (with half-lives of 3,000 yr, 10,000 yr and 50,000 yr, respectively). Most were first manufactured industrially in the 1950s, and many are measurable in firn air[102], and with large enough samples could be measured in ice cores[15]. But although they are measurable, distinct peaks are very recent and sometimes absent because major declines in industrial production are occurring after the negotiation and ratification of the 1989 Montreal and 2005 Kyoto protocols.

Of the various possible mid- to late-twentieth-century markers of the Great Acceleration, the global ^{14}C peak provides an unambiguously global change in a number of stratigraphic deposits. We suggest that an unequivocally annual record is the optimal choice to reflect the ^{14}C peak, thereby giving a dating accuracy of one year. We propose that the GSSP marker should be the ^{14}C peak, at 1964, within dated annual rings of a pine tree (*Pinus sylvestris*) from King Castle, Niepołomice, 25 km east of Kraków, Poland[103] (Fig. 2). Secondary correlated markers would include plutonium isotope ratios ($^{240}Pu/^{239}Pu$) in sediments indicating bomb testing[104], (fast-decaying) 137-Caesium[97], alongside the presence of peaks in very long-lived iodine isotopes (^{129}I, with half-life 15.7 million years) found in marine sediments[105] and soils[106].

While radionuclide fallout did not have major biological or other widespread physical repercussions, other auxiliary stratotypes may include the numerous other human-driven changes resulting in mid- to late-twentieth-century changes in geological deposits, including fossil pollen of novel genetically modified crops; declines in $\delta^{15}N$ in Northern Hemisphere

lakes[96] and ice cores[15]; the emergence of SF_6 and CF_4 from background levels[15]; lead isotopes in ice cores[15]; microplastics in marine sediments[97]; diatom assemblages in lakes in response to eutrophication[41]; and benthic foraminifera changes in marine sediments[41].

Dating the Anthropocene

We conclude that most proposed Anthropocene start dates, including the earliest detectable human impacts[42], earliest widespread impacts[45], and historic events such as the Industrial Revolution[1–3,39,40], can probably be rejected because they are not derived from a globally synchronous marker. Our review highlights that only those environmental changes associated with well-mixed atmospheric gases provide clearly global synchronous geological markers on an annual or decadal scale, as is required to define a GSSP for the Anthropocene. The earliest potential GSSP primary marker we identify is the inflection of atmospheric methane at 5,020 yr BP (Fig. 2; Table 1), but correlated auxiliary stratotypes are lacking. Thus, the CH_4 inflection is unlikely to be a strong candidate for the beginning of the Anthropocene. We find that only two other events—the Orbis spike dip in CO_2 with a minimum at 1610, and the bomb spike 1964 peak in ^{14}C—appear to fulfil the criteria for a GSSP to define the inception of the Anthropocene (Fig. 2; Table 1). While both GSSP dates have a number of correlated auxiliary stratotypes there are advantages and disadvantages associated with each.

The main advantage to the 1610 Orbis spike is the geological and historical importance of the event. In common with other epoch boundaries[10] this boundary would document changes in climate[87,107], chemistry[75] and palaeontological[65,85] signals. Critically, the transoceanic movement of species is an unambiguously permanent change to the Earth system[40], and such a boundary would mark Earth's last globally synchronous cool period[87] before the long-term global warmth of the Anthropocene Epoch. Historically, the Industrial Revolution has often been considered as the most important event in relation to the inception of the Anthropocene[1,2,39,40], but we have not identified a clear global Industrial Revolution GSSP. However, in the view of many historians, industrialization and extensive fossil fuel use were only made possible by the annexing of the Americas[88]. Before the Industrial Revolution both northwest Europe and southern China were similar in terms of life expectancy and material consumption patterns, including modest coal use, and both regions faced productive boundaries based on the available land area[88]. Thus, the agricultural commodities from the vast new lands of the Americas allowed Europe to transcend its ecological limits and sustain economic growth. In turn, this freed labour, allowing Europe to industrialize. That is, the Americas made industrialization possible owing to the unprecedented inflow of new cheap resources (and profitable new markets for manufactured goods). This 'Great Divergence' of Europe from the rest of the world required access to and exploitation of new lands plus a rich source of easily exploitable energy: coal[88]. Thus, dating the Anthropocene to start about 150 years before the beginning of the Industrial Revolution is consistent with a contemporary understanding of the likely material causes of the Industrial Revolution. The main disadvantage to the Orbis hypothesis is that a number of deposits may not show large changes around 1600, particularly in terms of biological material from the transport of species to new continents or oceans, because there are time-lags before species newly appear in geological deposits.

The key advantage of selecting 1964 as the base of a new Anthropocene Epoch is the sheer variety of human impacts recorded during the Great Acceleration: almost all stratigraphic records today, and over recent decades, have some marker of human activity. The latter part of the twentieth century is unambiguously a time of major anthropogenic global environmental impacts[108]. One disadvantage is that although nuclear explosions have the capacity to fundamentally transform many aspects of Earth's functioning, so far they have not done so, making the radionuclide spike a good GSSP marker but not an Earth-changing event. A further possible limitation in selecting such a recent date is that some deposits, notably some marine sediments, do not accumulate and stabilize over time spans as short as the past 50 years, making clear datable changes and correlation among some stratotypes sometimes difficult to discern[40].

Choosing between the 1610 Orbis and 1964 bomb spikes is challenging. As an alternative, a GSSA date, based on stratigraphic evidence, could be agreed upon by committee as the inception of the Anthropocene. However, any chosen date would be potentially open to challenge as arbitrary. For example, the Industrial Revolution is certainly a pivotal moment in human history, yet it is unclear how one could choose, based on the available geological evidence, an early Industrial Revolution GSSA date, say 1800, over a later date, perhaps 1850 or 1900. Similarly, the Great Acceleration is diachronous[108], and GSSA suggested dates could be 1945, 1950 or 1954 (ref. 109). Given such difficulties, given that GSSP markers are preferred[10], and given that candidate GSSP markers exist, a GSSA date seems unnecessary. Of the GSSP possibilities we tend to prefer 1610, because the transoceanic movement of species is a clear and permanent geological change to the Earth system. This date also fits more closely with Crutzen and Stoermer's original proposal[1] of an important historical juncture—the Industrial Revolution—as the beginning of the Anthropocene, which has been enduringly popular and useful, suggesting 1610 may be similarly so.

We hope that identifying a limited number of possible events and GSSP markers may assist in focusing research efforts to select a robust GSSP alongside a series of auxiliary stratotypes. Such research might include compiling data sets of the first appearance of non-native species in lake and marine sediments to better document the transoceanic spread of species and improve the evidence on which the 1610 proposal is based. The reliable detection of ^{129}I in high-resolution glacier ice and expanding the number of locations at which novel minerals, compounds and other recent human signals are investigated[41,47] would advance the 1964 GSSP proposal.

Ratification of an Anthropocene Epoch would require a further decision to be made, that is, whether to retain the Holocene Epoch (Fig. 1). All Anthropocene GSSP choices would leave a complete Holocene Epoch at least three orders of magnitude shorter than any other epoch[10] and similar to previous Pleistocene interglacials[55], which are not epoch-level events. Furthermore, the existence of a Holocene Epoch is due, in part, to the view—originating from nineteenth-century geologists—that the presence or influence of humans distinguished the Holocene from the Pleistocene[9,26,27,35,38]. An Anthropocene Epoch, combined with today's evidence that *Homo sapiens* is a Pleistocene species, removes key justifications for retaining the Holocene as an epoch-level designation. We therefore suggest that if the Anthropocene is accepted as an epoch it should directly follow the Pleistocene (Fig. 1c), as suggested independently elsewhere[110]. If the Holocene ceases to be an epoch but refers instead to the final stage of the Pleistocene Epoch, we suggest that the term Holocenian Stage is used, to maintain consistency with current terminology. While an alternative informal geological term, the Flandrian stage, denotes the current interglacial as part of the Pleistocene, its use has strongly declined over recent decades[10], and would not be as recognizable as the Holocenian Stage. Re-classifying any pre-Anthropocene Epoch interglacial time unit as the Holocenian Stage will create the usual tension[10] between resistance to altering past GTS agreements and the maintenance of GTS internal consistency.

The wider importance

The choice of either 1610 or 1964 as the beginning of the Anthropocene would probably affect the perception of human actions on the environment. The Orbis spike implies that colonialism, global trade and coal brought about the Anthropocene. Broadly, this highlights social concerns, particularly the unequal power relationships between different groups of people, economic growth, the impacts of globalized trade, and our current reliance on fossil fuels. The onward effects of the arrival of Europeans in the Americas also highlights a long-term and large-scale example of human actions unleashing processes that are difficult to predict or manage. Choosing the bomb spike tells a story of an elite-driven technological development that threatens planet-wide destruction. The

long-term advancement of technology deployed to kill people, from spears to nuclear weapons, highlights the more general problem of 'progress traps'[111]. Conversely, the 1963 Partial Test Ban Treaty and later agreements highlight the ability of people to collectively successfully manage a major global threat to humans and the environment. The event or date chosen as the inception of the Anthropocene will affect the stories people construct about the ongoing development of human societies.

Past scientific discoveries have tended to shift perceptions away from a view of humanity as occupying the centre of the Universe. In 1543 Copernicus's observation of the Earth revolving around the Sun demonstrated that this is not the case. The implications of Darwin's 1859 discoveries then established that *Homo sapiens* is simply part of the tree of life with no special origin. Adopting the Anthropocene may reverse this trend by asserting that humans are not passive observers of Earth's functioning. To a large extent the future of the only place where life is known to exist is being determined by the actions of humans. Yet, the power that humans wield is unlike any other force of nature, because it is reflexive and therefore can be used, withdrawn or modified. More widespread recognition that human actions are driving far-reaching changes to the life-supporting infrastructure of Earth may well have increasing philosophical, social, economic and political implications over the coming decades.

Received 26 March 2014; accepted 12 January 2015.

1. Crutzen, P. J. & Stoermer, E. F. The Anthropocene. *IGBP Global Change Newsl.* **41,** 17–18 (2000).
   This paper suggested that the Holocene has ended and the Anthropocene has begun, starting the contemporary increase in the usage of the term Anthropocene.
2. Crutzen, P. J. Geology of mankind. *Nature* **415,** 23 (2002).
3. Steffen, W., Crutzen, P. J. & McNeill, J. R. The Anthropocene: are humans now overwhelming the great forces of nature. *Ambio* **36,** 614–621 (2007).
4. Zalasiewicz, J., Williams, M., Haywood, A. & Ellis, M. The Anthropocene: a new epoch of geological time? *Phil. Trans. R. Soc. Lond. A* **369,** 835–841 (2011).
5. Dalby, S. Biopolitics and climate security in the Anthropocene. *Geoforum* **49,** 184–192 (2013).
6. Anon. The Anthropocene: a man-made world. *The Economist* **May 26** (2011); http://www.economist.com/node/18741749.
7. Zalasiewicz, J. *The Earth After Us: What Legacy Will Humans Leave in the Rocks*? (Oxford University Press, 2008).
8. Autin, W. J. & Holbrook, J. M. Is the Anthropocene an issue of stratigraphy or pop culture? *GSA Today* **22,** 60–61 (2012).
9. Gibbard, P. L. & Walker, M. J. C. The term 'Anthropocene' in the context of formal geological classification. *Geol. Soc. Lond. Spec. Publ.* **395,** 29–37 (2014).
   This paper presents a view that there is not currently enough evidence to formally ratify a new Anthropocene Epoch.
10. Gradstein, F. M., Ogg, J. G., Schmitz, M. D. & Ogg, G. M. *The Geologic Time Scale 2012* (Elsevier, 2012).
    This book is the latest GTS, including the formal assessments of Earth's history divided into epochs, periods, eras and eons.
11. Finney, S. C. The 'Anthropocene' as a ratified unit in the ICS International Chronostratigraphic Chart: fundamental issues that must be addressed by the Task Group. *Geol. Soc. Lond. Spec. Publ.* **395,** 23–28 (2014).
    This paper details the requirements and questions that will need to be addressed by the initial committee that will recommend whether or not an Anthropocene epoch is to be formally defined.
12. Canfield, D. E., Glazer, A. N. & Falkowski, P. G. The evolution and future of Earth's nitrogen cycle. *Science* **330,** 192–196 (2010).
13. Ciais, P. *et al.* in *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Stocker, T. F. *et al.*) Ch. 6, 465–570 (Cambridge Univ. Press, 2013).
14. Masson-Delmotte, V. *et al.* in *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Stocker, T. F. *et al.*) Ch. 5, 383–464 (Cambridge Univ. Press, 2013).
15. Wolff, E. W. Ice Sheets and the Anthropocene. *Geol. Soc. Lond. Spec. Publ.* **395,** 255–263 (2014).
16. International Geosphere-Biosphere Programme, Intergovernmental Oceanographic Commission, Scientific Committee on Oceanic Research. *Ocean Acidification Summary for Policymakers – Third Symposium on the Ocean in a High-CO_2 World* (International Geosphere-Biosphere Programme, 2013), http://ocean-acidification.net/for-policymakers/.
17. Running, S. W. A measurable planetary boundary for the biosphere. *Science* **337,** 1458–1459 (2012).
18. Krausmann, F. *et al.* Global human appropriation of net primary production doubled in the 20th century. *Proc. Natl Acad. Sci. USA* **110,** 10324–10329 (2013).
19. Barnosky, A. D. *et al.* Has the Earth's sixth mass extinction already arrived? *Nature* **471,** 51–57 (2011).
20. Thomas, C. D. The Anthropocene could raise biological diversity. *Nature* **502,** 7 (2013).
21. Baiser, B., Olden, J. D., Record, S., Lockwood, J. L. & McKinney, M. L. Pattern and process of biotic homogenization in the New Pangaea. *Proc. R. Soc. Lond. B* **279,** 4772–4777 (2012).
22. Palumbi, S. R. Humans as the world's greatest evolutionary force. *Science* **293,** 1786–1790 (2001).
23. Darimont, C. T. *et al.* Human predators outpace other agents of trait change in the wild. *Proc. Natl Acad. Sci. USA* **106,** 952–954 (2009).
24. Tabashnik, B. E., Mota-Sanchez, D., Whalon, M. E., Hollingworth, R. M. & Carriere, Y. Defining terms for proactive management of resistance to Bt crops and pesticides. *J. Econ. Entomol.* **107,** 496–507 (2014).
25. Stuart, Y. E. *et al.* Rapid evolution of a native species following invasion by a congener. *Science* **346,** 463–466 (2014).
26. Davis, R. V. Inventing the present: historical roots of the Anthropocene. *Earth Sci. Hist.* **30,** 63–84 (2011).
    This paper investigates and reviews the history of the use of the terms 'Holocene' and 'Anthropocene', showing that the Holocene includes humans in its first nineteenth-century definition.
27. Rudwick, M. S. J. *Bursting the Limits of Time: The Reconstruction of Geohistory in the Age of Revolution* (University of Chicago Press, 2005).
28. Jenkyn, T. W. Lessons in Geology XLVI. Chapter IV. On the effects of organic agents on the Earth's crust. *Popular Educator* **4,** 139–141 (1854).
29. Jenkyn, T. W. Lessons in Geology XLIX. Chapter V. On the classification of rocks section IV. On the tertiaries *Popular Educator* **4,** 312–316 (1854).
30. Hansen, P. H. *The Summits of Modern Man: Mountaineering after the Enlightenment* (Harvard University Press, 2013).
31. Haughton, S. *Manual of Geology* (Longman, 1865).
32. Stoppani, A. *Corso di Geologia* Vol. II (G. Bernardoni e G. Brigola, 1873).
33. Dana, J. D. *Manual of Geology* (Theodore Bliss and Co., 1863).
34. Le Conte, J. On critical periods in the history of the Earth and their relation to evolution; and on the Quaternary as such a period. *Am. J. Sci.* **14,** 99–114 (1877).
35. Lyell, C. *Principles of Geology* Volumes I, II and III (University of Chicago Press, 1990); originally published by John Murray, 1830–1833.
36. Shantser, E. V. in *Great Soviet Encyclopedia* Vol. 2 (ed. Prokhorov, A. M.) 139–144 (Macmillan, 1979).
37. Vernadsky, W. I. Biosphere and Noosphere. *Am. Sci.* **33,** 1–12 (1945).
38. Walker, M. *et al.* Formal definition and dating of the GSSP (Global Stratotype Section and Point) for the base of the Holocene using the Greenland NGRIP ice core, and selected auxiliary records. *J. Quat. Sci.* **24,** 3–17 (2009).
39. Steffen, W., Grinevald, J., Crutzen, P. & McNeill, J. The Anthropocene: conceptual and historical perspectives. *Phil. Trans. R. Soc. Lond. A* **369,** 842–867 (2011).
40. Zalasiewicz, J. *et al.* Stratigraphy of the Anthropocene. *Phil. Trans. R. Soc. Lond. A* **369,** 1036–1055 (2011).
41. Waters, C. N., Zalasiewicz, J. A., Williams, M., Ellis, M. A. & Snelling, A. M. A stratigraphical basis for the Anthropocene? *Geol. Soc. Lond. Spec. Publ.* **395,** 1–21 (2014).
    This paper reviews various stratigraphic markers relevant to defining the Anthropocene, with an up-to-date collation of the many markers coincident with the Industrial Revolution and the Great Acceleration.
42. Glikson, A. Fire and human evolution: the deep-time blueprints of the Anthropocene. *Anthropocene* **3,** 89–92 (2013).
43. Ruddiman, W. F. The Anthropocene. *Annu. Rev. Earth Planet. Sci.* **41,** 45–68 (2013).
    This paper summarizes the data and arguments that human activity altered CO_2 and CH_4 emissions thousands of years ago, leading to a delayed next glaciation, known as the Early Anthropogenic Hypothesis.
44. Foley, S. F. *et al.* The Palaeoanthropocene—the beginnings of anthropogenic environmental change. *Anthropocene* **3,** 83–88 (2013).
45. Balter, M. Archaeologists say the 'Anthropocene' is here—but it began long ago. *Science* **340,** 261–262 (2013).
46. Fischer-Kowalski, M., Krausmann, F. & Pallua, I. A sociometabolic reading of the Anthropocene: modes of subsistence, population size and human impact on Earth. *Anthropocene Rev.* **1,** 8–33 (2014).
    This paper takes an alternative view of the Anthropocene, considering human energy sources, and posits two transitions, to an agricultural mode, about 10,000 yr BP, and to an industrial mode, which begins after 1500.
47. Zalasiewicz, J., Williams, M. & Waters, C. N. Can an Anthropocene series be defined and recognized? *Geol. Soc. Lond. Spec. Publ.* **395,** 39–53 (2014).
48. Roebroeks, W. & Villa, P. On the earliest evidence for habitual use of fire in Europe. *Proc. Natl Acad. Sci. USA* **108,** 5209–5214 (2011).
49. Barnosky, A. D. Palaeontological evidence for defining the Anthropocene. *Geol. Soc. Lond. Spec. Publ.* **395,** 149–165 (2014).
50. Barnosky, A. D., Koch, P. L., Feranec, R. S., Wing, S. L. & Shabel, A. B. Assessing the causes of Late Pleistocene extinctions on the continents. *Science* **306,** 70–75 (2004).
51. Lorenzen, E. D. *et al.* Species-specific responses of Late Quaternary megafauna to climate and humans. *Nature* **479,** 359–364 (2011).
52. Ellis, E. C. *et al.* Used planet: a global history. *Proc. Natl Acad. Sci. USA* **110,** 7978–7985 (2013).
53. Certini, G. & Scalenghe, R. Anthropogenic soils are the golden spikes for the Anthropocene. *Holocene* **21,** 1269–1274 (2011).
54. Gale, S. J. & Hoare, P. G. The stratigraphic status of the Anthropocene. *Holocene* **22,** 1491–1494 (2012).

55. Tzedakis, P. C., Channell, J. E. T., Hodell, D. A., Kleiven, H. F. & Skinner, L. C. Determining the natural length of the current interglacial. *Nature Geosci.* **5,** 138–141 (2012).
56. Broecker, W. C. & Stocker, T. F. The Holocene CO_2 rise: Anthropogenic or natural? *Eos* **87,** 27–29 (2006).
57. Stocker, B. D., Strassmann, K. & Joos, F. Sensitivity of Holocene atmospheric CO_2 and the modern carbon budget to early human land use: analyses with a process-based model. *Biogeosciences* **8,** 69–88 (2011).
58. Kaplan, J. O. *et al.* Holocene carbon emissions as a result of anthropogenic land cover change. *Holocene* **21,** 775–791 (2011).
59. Blunier, T., Chappellaz, J., Schwander, J., Stauffer, B. & Raynaud, D. Variations in atmospheric methane concentration during the Holocene epoch. *Nature* **374,** 46–49 (1995).
60. Sapart, C. J. *et al.* Natural and anthropogenic variations in methane sources during the past two millennia. *Nature* **490,** 85–88 (2012).
61. Singarayer, J. S., Valdes, P. J., Friedlingstein, P., Nelson, S. & Beerling, D. J. Late Holocene methane rise caused by orbitally controlled increase in tropical sources. *Nature* **470,** 82–85 (2011).
62. Diamond, J. *Guns, Germs and Steel: A Short History of Everybody for the Last 13,000 Years* (Chatto and Windus, 1997).
63. Mann, C. C. *1493: How the Ecological Collision of Europe and the Americas Gave Rise to the Modern World* (Granta, 2011).
64. Crosby, A. W. *The Columbian Exchange: Biological and Cultural Consequences of 1492* 30 yr edn (Preager, 2003).
65. Mercuri, A. M. *et al.* A marine/terrestrial integration for mid-late Holocene vegetation history and the development of the cultural landscape in the Po valley as a result of human impact and climate change. *Vegetat. Hist. Archaeobot.* **21,** 353–372 (2012).
66. Piperno, D. R. Identifying crop plants with phytoliths (and starch grains) in Central and South America: a review and an update of the evidence. *Quat. Int.* **193,** 146–159 (2009).
67. Zalasiewicz, J. & Williams, M. The Anthropocene: a comparison with the Ordovician-Silurian boundary. *Rendiconti Lincei-Scienze Fisiche E Naturali* **25,** 5–12 (2014).
68. Denevan, W. M. *The Native Population of the Americas in 1492* 2nd edn (University of Wisconsin Press, 1992).
69. Mann, C. C. *1491: New Revelations of the Americas Before Columbus* (Vintage, 2005).
70. Nevle, R. J. & Bird, D. K. Effects of syn-pandemic fire reduction and reforestation in the tropical Americas on atmospheric CO_2 during European conquest. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **264,** 25–38 (2008).
This paper presents a synthesis of data computing the impacts of the rapid 1492–1650 reduction in population across the Americas and the carbon uptake implications.
71. Dull, R. A. *et al.* The Columbian encounter and the Little Ice Age: abrupt land use change, fire, and greenhouse forcing. *Ann. Assoc. Am. Geogr.* **100,** 755–771 (2010).
72. Nevle, R. J., Bird, D. K., Ruddiman, W. F. & Dull, R. A. Neotropical human-landscape interactions, fire, and atmospheric CO_2 during European conquest. *Holocene* **21,** 853–864 (2011).
73. Ahn, J. *et al.* Atmospheric CO_2 over the last 1000 years: a high-resolution record from the West Antarctic Ice Sheet (WAIS) divide ice core. *Glob. Biogeochem. Cycles* **26,** GB2027 (2012).
74. Rubino, M. *et al.* A revised 1000 year atmospheric delta C-13-CO_2 record from Law Dome and South Pole, Antarctica. *J. Geophys. Res. D* **118,** 8482–8499 (2013).
75. MacFarling Meure, C. *et al.* Law Dome CO_2, CH_4 and N_2O ice core records extended to 2000 years BP. *Geophys. Res. Lett.* **33,** L14810 (2006).
76. Etheridge, D. M., Steele, L. P., Francey, R. J. & Langenfelds, R. L. Atmospheric methane between 1000 AD and present: evidence of anthropogenic emissions and climatic variability. *J. Geophys. Res. D* **103,** 15979–15993 (1998).
77. Smith, V. C. Volcanic markers for dating the onset of the Anthropocene. *Geol. Soc. Lond. Spec. Publ.* **395,** 283–299 (2014).
78. de Silva, S. L. & Zielinski, G. A. Global influence of the AD1600 eruption of Huaynaputina, Peru. *Nature* **393,** 455–458 (1998).
79. Thompson, L. G. *et al.* Annually resolved ice core records of tropical climate variability over the past ~1800 Years. *Science* **340,** 945–950 (2013).
80. Power, M. J. *et al.* Climatic control of the biomass-burning decline in the Americas after AD 1500. *Holocene* **23,** 3–13 (2013).
81. Wang, Z., Chappellaz, J., Park, K. & Mak, J. E. Large variations in Southern Hemisphere biomass burning during the last 650 years. *Science* **330,** 1663–1666 (2010).
82. Ferretti, D. F. *et al.* Unexpected changes to the global methane budget over the past 2000 years. *Science* **309,** 1714–1717 (2005).
83. Mischler, J. A. *et al.* Carbon and hydrogen isotopic composition of methane over the last 1000 years. *Glob. Biogeochem. Cycles* **23,** GB4024 (2009).
84. Mitchell, L. E., Brook, E. J., Sowers, T., McConnell, J. R. & Taylor, K. Multidecadal variability of atmospheric methane, 1000–1800 CE. *J. Geophys. Res.* **116,** G02007 (2011).
85. Bush, M. B. & Colinvaux, P. A. Tropical forest disturbance: Paleoecological records from Darien, Panama. *Ecology* **75,** 1761–1768 (1994).
86. Kinnard, C. *et al.* Reconstructed changes in Arctic sea ice over the past 1,450 years. *Nature* **479,** 509–512 (2011).
87. Neukom, R. *et al.* Inter-hemispheric temperature variability over the past millennium. *Nature Clim. Change* **4,** 362–367 (2014).
This paper synthesizes paleoclimate records from the southern and northern hemispheres, showing one globally synchronous cool period (1594–1677) and one globally synchronous warm period (1965 onwards) within the last 1,000 years.
88. Pomeranz, K. *The Great Divergence: China, Europe, and the Making of the Modern World Economy* (Princeton University Press, 2000).
89. Wallerstein, I. *The Modern World-System I: Capitalist Agriculture and the Origins of the European World-Economy in the Sixteenth Century* (Academic Press, 1974).
90. Killick, D. & Fenn, T. Archaeometallurgy: the study of preindustrial mining and metallurgy. *Annu. Rev. Anthropol.* **41,** 559–575 (2012).
91. Cooke, C. A., Balcom, P. H., Biester, H. & Wolfe, A. P. Over three millennia of mercury pollution in the Peruvian Andes. *Proc. Natl Acad. Sci. USA* **106,** 8830–8834 (2009).
92. Hong, S. M., Candelone, J. P., Patterson, C. C. & Boutron, C. F. History of ancient copper smelting pollution during Roman and medieval times recorded in Greenland ice. *Science* **272,** 246–249 (1996).
93. Rose, N. L. & Appleby, P. G. Regional applications of lake sediment dating by spheroidal carbonaceous particle analysis I: United Kingdom. *J. Paleolimnol.* **34,** 349–361 (2005).
94. Snowball, I., Hounslow, M. W. & Nilsson, A. Geomagnetic and mineral magnetic characterization of the Anthropocene. *Geol. Soc. Lond. Spec. Publ.* **395,** 119–141 (2014).
95. Wolfe, A. P. *et al.* Stratigraphic expressions of the Holocene-Anthropocene transition revealed in sediments from remote lakes. *Earth Sci. Rev.* **116,** 17–34 (2013).
96. Holtgrieve, G. W. *et al.* A coherent signature of Anthropogenic nitrogen deposition to remote watersheds of the Northern Hemisphere. *Science* **334,** 1545–1548 (2011).
97. Gałuszka, A., Migaszewski, Z. M. & Zalasiewicz, J. Assessing the Anthropocene with geochemical methods. *Geol. Soc. Lond. Spec. Publ.* **395,** 221–238 (2014).
98. Falkowski, P. *et al.* The global carbon cycle: a test of our knowledge of Earth as a system. *Science* **290,** 291–296 (2000).
99. Fairchild, I. J. & Frisia, S. Definition of the Anthropocene: a view from the underworld. *Geol. Soc. Lond. Spec. Publ.* **395,** 239–254 (2014).
100. Hua, Q. Radiocarbon: a chronological tool for the recent past. *Quat. Geochronol.* **4,** 378–390 (2009).
101. Harnisch, J. & Eisenhauer, A. Natural CF_4 and SF_6 on Earth. *Geophys. Res. Lett.* **25,** 2401–2404 (1998).
102. Butler, J. H. *et al.* A record of atmospheric halocarbons during the twentieth century from polar firn air. *Nature* **399,** 749–755 (1999).
103. Rakowski, A. Z. *et al.* Radiocarbon method in environmental monitoring of CO_2 emission. *Nucl. Instrum. Methods Phys. Res. B* **294,** 503–507 (2013).
104. Ketterer, M. E. *et al.* Resolving global versus local/regional Pu sources in the environment using sector ICP-MS. *J. Anal. At. Spectrom.* **19,** 241–245 (2004).
105. Fehn, U. *et al.* Determination of natural and anthropogenic I-129 in marine sediments. *Geophys. Res. Lett.* **13,** 137–139 (1986).
106. Hansen, V., Roos, P., Aldahan, A., Hou, X. & Possnert, G. Partition of iodine (I-129 and I-127) isotopes in soils and marine sediments. *J. Environ. Radioact.* **102,** 1096–1104 (2011).
107. Schurer, A. P., Hegerl, G. C., Mann, M. E., Tett, S. F. B. & Phipps, S. J. Separating forced from chaotic climate variability over the past millennium. *J. Clim.* **26,** 6954–6973 (2013).
108. Steffen, W., Broadgate, W., Deutsch, L., Gaffney, O. & Ludwig, C. The trajectory of the Anthropocene: the Great Acceleration. *Anthropocene Rev.* http://dx.doi.org/10.1177/2053019614564785 (in the press).
109. Zalasiewicz, J. *et al.* When did the Anthropocene begin? A mid-twentieth century boundary level is stratigraphically optimal. *Quat. Int.* http://dx.doi.org/10.1016/j.quaint.2014.11.045 (in the press).
110. van der Pluijm, B. Hello Anthropocene, goodbye Holocene. *Earth's Future* **2,** 2014EF000268 (2014).
111. Wright, R. *A Short History of Progress* (House of Anansi Press, 2004).
112. Shakun, J. D. *et al.* Global warming preceded by increasing carbon dioxide concentrations during the last deglaciation. *Nature* **484,** 49–54 (2012).
113. Monnin, E, *et al.* Atmospheric CO_2 concentrations over the last glacial termination. *Science* **291,** 112–114 (2001).
114. Veres, D. *et al.* The Antarctic ice core chronology (AICC2012): an optimized multi-parameter and multi-site dating approach for the last 120 thousand years. *Clim. Past* **9,** 1733–1748 (2013).
115. Marcott, S. A., Shakun, J. D., Clark, P. U. & Mix, A. C. A reconstruction of regional and global temperature for the past 11,300 years. *Science* **339,** 1198–1201 (2013).
116. Alexander, L. V, *et al.* in *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Stocker, T. F. *et al.*) 3–28 (Cambridge Univ. Press, 2013).
117. Indermuhle, A. *et al.* Holocene carbon-cycle dynamics based on CO_2 trapped in ice at Taylor Dome, Antarctica. *Nature* **398,** 121–126 (1999).
118. Siegenthaler, U. *et al.* Supporting evidence from the EPICA Dronning Maud Land ice core for atmospheric CO_2 changes during the past millennium. *Tellus B* **57,** 51–57 (2005).
119. Ahn, J. *et al.* CO_2 diffusion in polar ice: observations from naturally formed CO_2 spikes in the Siple Dome (Antarctica) ice core. *J. Glaciol.* **54,** 685–695 (2008).
120. Marín-Spiotta, E. & Sharma, S. Carbon storage in successional and plantation forest soils: a tropical analysis. *Glob. Ecol. Biogeogr.* **22,** 105–117 (2013).

121. Bonner, M. T. L., Schmidt, S. & Shoo, L. P. A meta-analytical global comparison of aboveground biomass accumulation between tropical secondary forests and monoculture plantations. *For. Ecol. Manage.* **291,** 73–86 (2013).
122. Pongratz, J., Caldeira, K., Reick, C. H. & Claussen, M. Coupled climate-carbon simulations indicate minor global effects of wars and epidemics on atmospheric CO_2 between AD 800 and 1850. *Holocene* **21,** 843–851 (2011).
123. Orihuela-Belmonte, D. E. *et al.* Carbon stocks and accumulation rates in tropical secondary forests at the scale of community, landscape and forest type. *Agric. Ecosyst. Environ.* **171,** 72–84 (2013).
124. Francey, R. J. *et al.* A 1000-year high precision record of $\delta^{13}C$ in atmospheric CO_2. *Tellus B* **51,** 170–193 (1999).
125. Trudinger, C. M., Enting, I. G., Francey, R. J., Etheridge, D. M. & Rayner, P. J. Long-term variability in the global carbon cycle inferred from a high-precision CO_2 and $\delta^{13}C$ ice-core record. *Tellus B* **51,** 233–248 (1999).
126. Böhm, F. *et al.* Evidence for preindustrial variations in the marine surface water carbonate system from coralline sponges. *Geochem. Geophys. Geosyst.* **3,** 1–13 (2002).
127. Trudinger, C. M., Enting, I. G., Rayner, P. J. & Francey, R. J. Kalman filter analysis of ice core data—2. Double deconvolution of CO_2 and $\delta^{13}C$ measurements. *J. Geophys. Res. D* **107,** D20 (2002).

Acknowledgements We acknowledge C. Hamilton, whose idea that humans are a reflexive power rather than force of nature was presented at the 'Thinking the Anthropocene' conference in Paris on 15 November 2013, and used with permission. We thank J. Kaplan and K. Krumhardt for the estimates of the population of the Americas, M. Irving for assistance with the figures, and C. Brierley, M.-E. Carr, W. Laurance, A. Mackay, O. Morton, R. Newman and C. Tzedakis for constructive discussion and remarks, and reviewer P. Gibbard for important comments. This work was funded by the European Research Council (T-FORCES, S.L.L.), a Philip Leverhulme Prize award (S.L.L.), and a Royal Society Wolfson Research Merit Award (M.A.M.).

Author Contributions S.L.L. and M.A.M. conceived the paper structure. S.L.L. conceived and developed the Obris hypothesis. S.L.L. wrote the geological importance, historical, farming and Orbis evidence reviews. M.A.M. wrote the Pleistocene, and industrialization and Great Acceleration evidence reviews. M.A.M. conceived and developed the figures. The final two sections, written by S.L.L., emerged from discussions between S.L.L. and M.A.M.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.L.L. (s.l.lewis@ucl.ac.uk).

ARTICLE

doi:10.1038/nature14279

Quantitative evolutionary dynamics using high-resolution lineage tracking

Sasha F. Levy[1,2,3]*, Jamie R. Blundell[4,5]*, Sandeep Venkataram[5], Dmitri A. Petrov[5], Daniel S. Fisher[4,5] & Gavin Sherlock[1]

Evolution of large asexual cell populations underlies ~30% of deaths worldwide, including those caused by bacteria, fungi, parasites, and cancer. However, the dynamics underlying these evolutionary processes remain poorly understood because they involve many competing beneficial lineages, most of which never rise above extremely low frequencies in the population. To observe these normally hidden evolutionary dynamics, we constructed a sequencing-based ultra high-resolution lineage tracking system in *Saccharomyces cerevisiae* that allowed us to monitor the relative frequencies of ~500,000 lineages simultaneously. In contrast to some expectations, we found that the spectrum of fitness effects of beneficial mutations is neither exponential nor monotonic. Early adaptation is a predictable consequence of this spectrum and is strikingly reproducible, but the initial small-effect mutations are soon outcompeted by rarer large-effect mutations that result in variability between replicates. These results suggest that early evolutionary dynamics may be deterministic for a period of time before stochastic effects become important.

A major focus of biomedical research has been to identify mutations responsible for increased pathogenicity, cancer progression, or drug resistance in large evolving asexual cell populations[1–12]. Yet, even characterizing all mutations underlying a disease is not sufficient to understand its progression. Rather, a quantitative understanding of the evolutionary dynamics is necessary to determine which adaptive mutations contribute significantly to driving the population fitness higher, and which are serendipitous or inconsequential. Mutations identified through genome sequencing are likely to constitute only the 'tip of the iceberg', with many beneficial mutations that impact the evolutionary dynamics never rising above extremely low frequencies[13,14].

A lineage trajectory, the size of a small subpopulation of cells over time, can be used to discover a beneficial mutation present at an extremely low frequency, and to measure its time of occurrence and selective advantage (Fig. 1a)[1,15–18]. A lineage increasing in size faster than can be explained by stochastic drift indicates that a beneficial mutation has occurred and risen to a high enough frequency to grow almost deterministically (that is, it has 'established'). Most beneficial mutations will drift to extinction before establishing (Supplementary Information section 4.1 and 4.4). For those that do establish, the exponential rate at which a lineage grows is a measure of the fitness effect (s) of the mutation. Extrapolating back the exponential growth, the establishment time (τ) can be inferred: this is a rough estimate of when the mutation occurred[19] (Supplementary Information section 4.1 and 4.2). A systematic characterization of the distributions of s and τ for beneficial mutations has been lacking, although these are fundamental to the evolutionary dynamics of large populations[20].

The major experimental challenge is developing a method to quantitatively measure the trajectories of large numbers of small lineages. Large lineages will accumulate multiple beneficial mutations contemporaneously, confounding measurements of s and τ (Fig. 1a, multiple mutations, Supplementary Information section 4.5). Small lineages are unlikely to acquire a beneficial mutation at all, so many trajectories must be observed to characterize the distributions of s and τ. DNA barcodes offer a powerful way to simultaneously track multiple lineages[21–23], yet technical barriers have limited the number of barcodes that can be inserted into cells[24]. Here we constructed a system capable of inserting ~500,000 random DNA barcodes into an initially clonal yeast population. Using this system in populations of $\sim 10^8$ cells growing in a defined glucose-limited minimal medium, we identified ~25,000 lineages that gained a beneficial mutation within ~168 generations, measured s and τ for each, and determined the spectrum of mutation rates to each fitness effect. This spectrum results in a deterministic increase in the mean population fitness early, with stochastic events governing its trajectory later.

Lineage tracking with random barcodes

We generated yeast lineages by inserting a random 20-nucleotide barcode at a single location in the genome (Fig. 1b, Supplementary Information section 1.3). To achieve a large number of integration events, we inserted a 'landing pad' into a neutral location in the yeast genome that allows for high-frequency, site-specific genomic integration of plasmids via the Cre–*loxP* recombination system[25,26]. A plasmid library containing ~500,000 random barcodes was inserted into the genome at the landing pad. Barcoding requires ~48 generations of growth from a common ancestor (Extended Data Fig. 1). Adaptive mutations begin to occur during this initial growth and can be carried forward into the evolution.

The same barcoded yeast library was evolved in replicate experiments (E1 and E2) for ~168 generations in serial batch culture, diluting 1:250 every ~8 generations, with a bottleneck population size of $\sim 7 \times 10^7$ (Extended Data Fig. 1, Supplementary Information section 4.4). To count the relative frequency of each lineage across time, we isolated genomic DNA from the pooled population, amplified lineage tags using a two-step PCR protocol, and sequenced amplicons (Fig. 1b, Supplementary Information section 1.5 and 5.2).

Plotting the relative frequency of each barcode over ~168 generations shows a reproducible pattern of population dynamics across replicates (Fig. 2a and Extended Data Fig. 2a). Most lineages declined in frequency (blue lines, neutral lineages), but a modest fraction (~5%,

[1]Department of Genetics, Stanford University, Stanford, California 94305-5120, USA. [2]Laufer Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, New York 11794–5252, USA. [3]Department of Biochemistry and Cellular Biology, Stony Brook University, Stony Brook, New York 11794–5215, USA. [4]Department of Applied Physics, Stanford University, Stanford, California 94305, USA. [5]Department of Biology, Stanford University, Stanford, California 94305, USA.
*These authors contributed equally to this work.

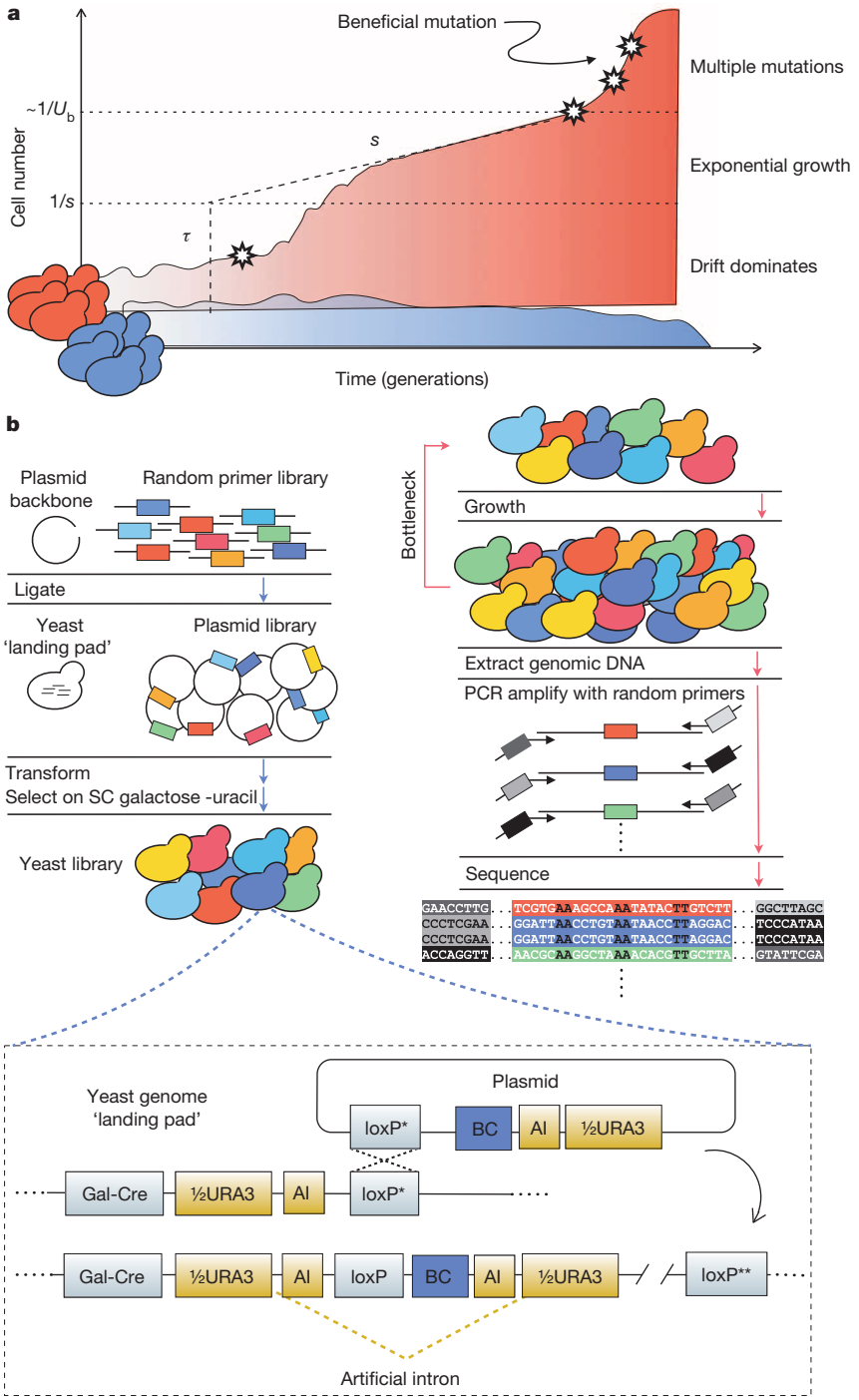


Figure 1 | Lineage tracking with random barcodes. **a**, Typical lineage trajectories. A small lineage that does not acquire a beneficial mutation (neutral, blue) will fluctuate in size due to drift before eventually being outcompeted. Rarely, a lineage will acquire a beneficial mutation (star) with a fitness effect of s (adaptive, red). In most cases, this beneficial mutation is lost to drift. If the beneficial mutants drift to a size $>\sim 1/s$ (lower dotted horizontal line), the lineage will begin to grow exponentially at a rate s. Extrapolating the exponential growth to the time at which the mutation is inferred to have reach a size $\sim 1/s$ yields the establishment time (τ, dashed vertical line) which roughly corresponds to the time when the mutation occurred with an uncertainty of $\sim 1/s$. At sizes $> \sim 1/U_b$ (upper dotted horizontal line), where U_b is the total beneficial mutation rate, the lineage will acquire additional beneficial mutations. **b**, Barcode insertion and sequencing. Left, sequences containing random 20 nucleotide barcodes (colours) are inserted first into a plasmid and then into a specific location in the genome. Bottom, recombination between two partially crippled *loxP* sites (*loxP**) integrates the plasmid into the genome and completes a *URA3* selectable marker, resulting in one functional and one crippled *loxP* site (*loxP***). The *URA3* marker is interrupted by an artificial intron containing the barcode. Right, to measure relative fitness, cells are passed through growth-bottleneck cycles of ~8 generations. Before each bottleneck, genomic DNA is extracted, lineage barcode tags are amplified using a two-step PCR protocol, and amplicons are sequenced. By inserting unique molecular identifiers[49] (also short random barcodes, grey bars) in early cycles of the PCR, PCR duplicates of the same template molecule (purple) are detected[49,50].

see below) had acquired a beneficial mutation that established (red lines, adaptive lineages). At later time points, the growth of adaptive lineages attenuates as the population mean fitness increases (clonal interference)[27].

To calculate the probability that a lineage contains an adaptive mutation, one must differentiate between a trajectory that increases due to an adaptive event from one that increases due to genetic drift and measurement errors. Because either scenario is rare, the right-hand tail of the distribution of read numbers is particularly important. Thus, we characterized the full distribution of noise that results from drift and sampling errors due to DNA isolation, amplification and sequencing, (black curve, Extended Data Fig. 2b, Supplementary Information section 5). The decline in frequency of neutral lineages is used to infer the increase in mean fitness of the population[4] (Fig. 2a and Extended Data Fig. 2a, b).

Using our estimates of noise and the mean fitness, we calculate the probability that a trajectory is explained by a mutation with fitness effect, s, having an establishment time, τ, over a broad range of s and τ (under a uniform prior in τ and an exponential prior in s, Extended Data 2c). If this exceeds the probability that no beneficial mutation occurs, we define the lineage as adaptive, with the peak of the probability our best estimate of s and τ (Supplementary Information section 7). Estimates of s and τ for each adaptive lineage are combined to calculate a second measurement of the increase in mean fitness (Fig. 2a and Extended Data Fig. 2a, insets). Our two methods to infer mean fitness agree, indicating that most lineages driving the mean fitness have been detected. Uncertainties in s and τ depend on the specific lineage trajectory; however, they are generally low ($\Delta s \pm 0.5\%$, $\Delta\tau \pm 10$ generations, Supplementary Information section 7.7).

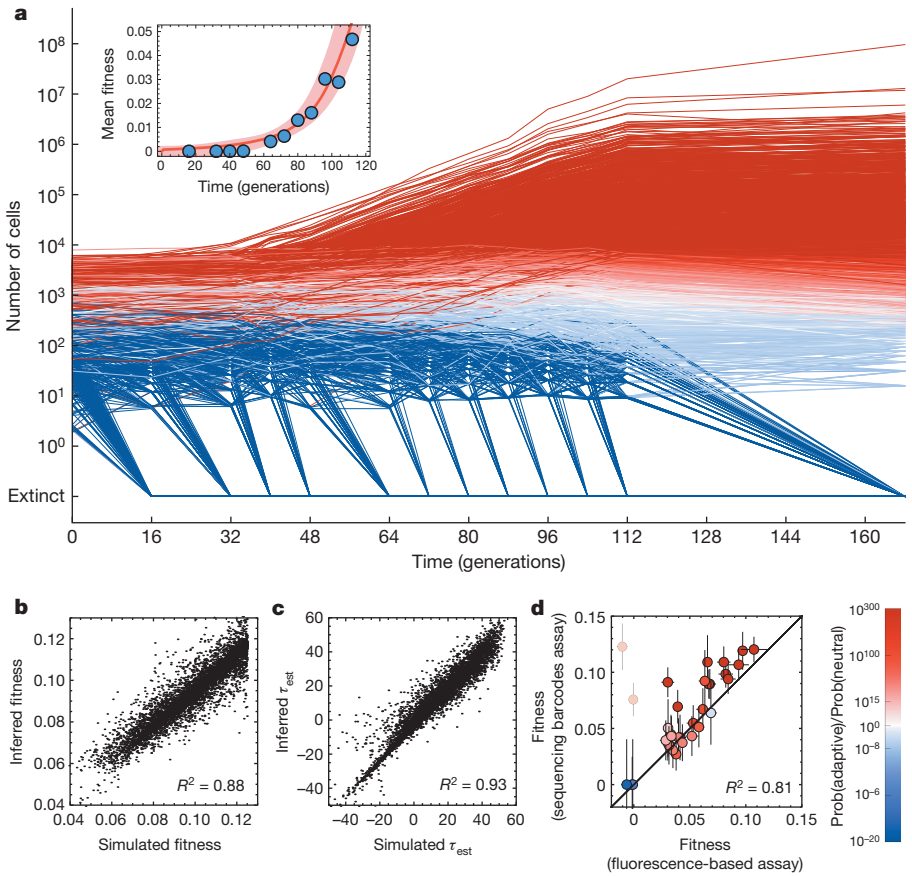


Figure 2 | Inferring the fitnesses and establishment times from lineage trajectories. **a**, Selected lineage trajectories from E1 coloured according to the probability that they contain an established beneficial mutation. The decline of adaptive lineages at later times is caused by the increase of the population mean fitness (inset). The population mean fitness is inferred from both the decline of neutral lineages (blue circles) and the growth of beneficial lineages (red line, Supplementary Information section 6.2). Shading indicates the error in mean fitness. **b**, **c**, The inferred fitnesses (**b**) and establishment times (**c**) from analysis of simulated trajectories correlate strongly with the known simulated values. **d**, Scatter plot of the fitness of 33 clones picked from E2 at generation 88 inferred by sequencing and pairwise competition (colouring as in (**a**), with outliers lightened in colour and excluded from correlation). Error bars represent one standard deviation.

To validate estimates of s and τ, we first analysed a simulated data set with comparable levels of noise to our experiment (Supplementary Information section 12). We find a strong correlation between the known and inferred values for both s ($R^2 = 0.88$ in Fig. 2b) and τ ($R^2 = 0.93$ in Fig. 2c). Second, we picked 33 clones from generation 88 that belong to different adaptive lineages and performed pairwise competitive fitness assays on each (Supplementary Information section 2). We find a strong correlation between these two methods ($R^2 = 0.81$, Fig. 2d). Outliers (lighter coloured data points) are likely due to a neutral cell being sampled from a lineage containing mostly adaptive cells. Other deviations could be due to interactions between adaptive lineages (that is, frequency dependent fitness) or to multiple mutations on the same genome (Supplementary Information section 8).

In total, ~25,000 beneficial mutations with a fitness effect of >2% established before generation 112 in E1 (Fig. 3a), a number that is roughly consistent with E2 (Extended Data Fig. 3a) and simulated data (Supplementary Figs 44 and 45 and Supplementary Information section 12). Adaptation occurs quickly: by generation 112 the population mean fitness is over 5% higher than the ancestor, with some lineages having a fitness advantage of >10%. E1 and E2 share 48 generations of common growth. During this time, ~6,000 lineages acquire a beneficial mutation that is sampled into, and establishes in, both replicates (Fig. 3a and Extended Data Fig. 3a, purple circles). We define these mutations as 'pre-existing': their presence is not an artefact of our experiment, but a general expectation for large populations grown from a single cell.

Beneficial mutation rates

To estimate the spectrum of beneficial mutation rates in the serial batch conditions, we consider only lineages that are identified as adaptive in one replicate but not the other (that is, are unlikely to contain mutations that occurred before barcoding, Supplementary Information section 9 and 10). Analysing the total number of cells with each s yields the best estimate of the mutation rate spectrum (Fig. 3a and Extended Data Fig. 3a, insets, and Supplementary Information section 11.1). These estimates are worse for fitness effects that have only occurred a few times. We find that beneficial mutations with $s > 5\%$ occur at a rate of $\sim 1 \times 10^{-6}$ per cell per generation (Supplementary Information section 11.2, Fig. 3a and Extended Data Fig. 3a, insets), a rate that is consistent across replicates. Using a fluctuation test[28,29], we find that the ancestor to our barcoded strains has a spontaneous mutation rate in non-repeat regions of $\sim 4 \times 10^{-10}$ per nucleotide per generation (Supplementary Information section 1.7)[30,31]. This implies that mutations in ~0.04% of the genome, ~5,000 bases, confer beneficial fitness effects of >5%. This target size is broadly consistent with previous reports[32,33], although it will certainly depend on the selective conditions. The beneficial mutation rate includes all events that have a heritable effect on fitness, and could include point mutations, indels, large genomic rearrangements or duplications, whole-genome duplications, and possibly even heritable epigenetic modifications. Reported beneficial mutation rates depend on the range of fitness effects that can establish and be detected. For example, if we include lower fitness effect mutations that are mostly pre-existing ($2\% < s < 5\%$), we find a beneficial mutation rate that is ~50× higher. However, as we discuss below, the total beneficial mutation rate is not necessary for a predictive understanding of the evolutionary dynamics. Instead, knowledge of the rate of mutation to the range of fitness effects that drive the dynamics is what is needed.

Mutation rate spectrum

Several authors have used extreme value theory to predict that the spectrum of beneficial mutation rates is exponential[34,35], with some experiments that sample small numbers of beneficial mutations supporting[17,36,37] or contradicting[38] this prediction. We do not find support for an exponential or even a monotonically decreasing distribution. Rather, most mutations we observe are confined to a narrow range of fitness effects ($2\% < s < 5\%$). At larger fitnesses, the distribution is relatively flat with two slight peaks in the fitness ranges 7–8% and 10–11%, a feature that is

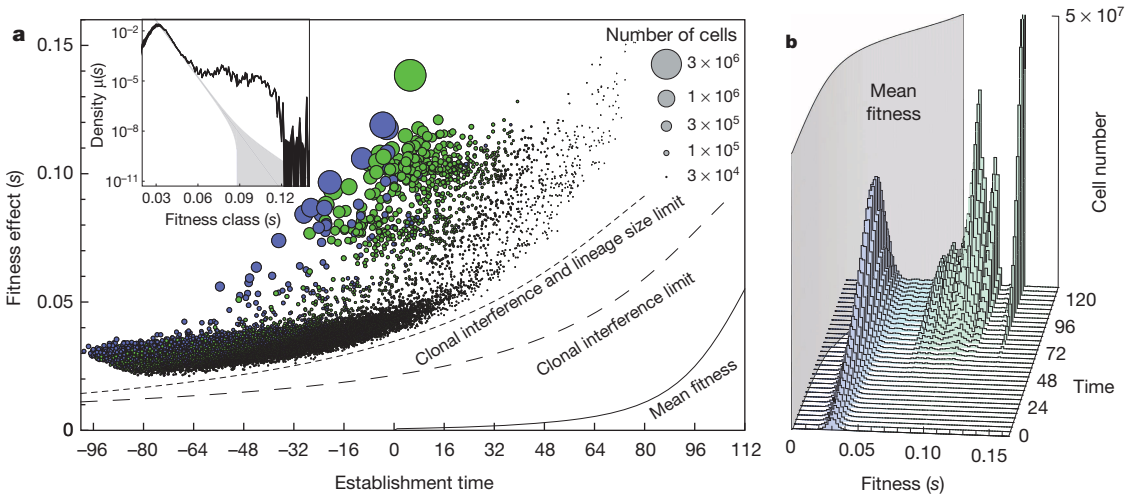


Figure 3 | Fitness effects, establishment times, and population dynamics. **a**, Scatter plot of τ and s of all ~25,000 beneficial mutations (circles) identified in E1. Circle area represents the size of the lineage at generation 88. Purple circles indicate lineages with mutations that occurred in the period of common growth ($t < 0$) that were sampled into, and established in, E1 and E2. Green circles indicate lineages that were identified as adaptive in only one replicate and likely contain mutations that arose after $t = 0$. Lines indicate the time limits before which mutations must occur in order to establish (large dash) or be observed (small dash). These limits trail the mean fitness (solid line) by ~$1/s$ generations. Inset, the spectrum of mutation rates, $\mu(s)$, as a function of fitness effect, s inferred from mutations that likely occurred after $t = 0$ (Supplementary Information section 10.2). The y axis is the mutation rate density, so the mutation rate to a range, Δs, is obtained by multiplying this density by Δs. The total beneficial mutation rate to $s > 5\%$ is inferred to be $\sim 1 \times 10^{-6}$ and is consistent across replicates. The observed spectrum is not exponential (grey line, with the error range shaded). **b**, the distribution of the number of adaptive cells binned by their fitness over time. As the mean fitness (grey curtain) surpasses the fitness of a subpopulation, cells with that fitness begin to decline in frequency.

consistent across replicates (Fig. 3a and Extended Data Fig. 3a, insets). Mutation rates to these two peaks are consistent with genomic target sizes of loss-of-function mutations for a single gene (~300 base pairs[31]); these have previously been shown to be adaptive in yeast grown in simple environments[1,3,4,39]. Weaker effect mutations ($s < 2\%$), which are hard to detect because they are rapidly outcompeted before establishing, do not occur at high enough rates to impact the population dynamics (Supplementary Information section 9.3).

Distribution of establishment times

For mutations that establish, τ roughly corresponds to the time at which a beneficial mutation occurred, with an uncertainty of a few times $1/s$ due to variability in initial stochastic drift (Supplementary Information section 4.1). Establishment times are broadly distributed ($-90 < \tau < 48$). Lineages containing beneficial mutations with very negative τ ($-90 < \tau < -40$) are usually identified as adaptive in both replicates (Fig. 3a and Extended Data Fig. 3a, purple). Establishment times as negative as -90 generations are expected[39] because of beneficial mutations that occur during the period of common growth ($t < 0$, Supplementary Information section 10.1). Indeed the number of pre-existing beneficial mutations is broadly consistent with the beneficial mutation rate we infer (Supplementary Information section 10). We observe very few mutations with $\tau > 48$ for the reasons that follow.

A beneficial mutation with a fitness effect s, that occurs in generation t will typically take another ~ $1/s$ generations to reach a size large enough to grow exponentially[19]. If before this time the mean fitness has increased by more than s, the mutation will decline in frequency and never grow exponentially. Thus, there is a time limit after which a beneficial mutation that occurs is unlikely to establish (Fig. 3a and Extended Data Fig. 3a, larger dashed lines). This time limit is shorter for smaller s for two reasons: (1) small s mutations must drift to higher numbers in order to establish, and (2) the mean fitness of the population surpasses its fitness advantage in a shorter time. A mutation with $s < 2\%$ is therefore extremely unlikely to establish because this limit is reached quickly. Thus, a fundamental lower limit on which fitness-effects can establish emerges from the population dynamics.

In addition to establishing, beneficial mutations in our assay must also grow to a large enough number to be detectable above the number of neutral (ancestral) cells remaining in its lineage. This shortens the time window in which a beneficial mutation must occur to be observed (Fig. 3a and Extended Data Fig. 3a, smaller dashed lines and Supplementary Information section 9). Beneficial mutations we are unable to detect (those occurring close to, or after, the time limit) never reach sizes much above their establishment number ($1/s$), are rapidly outcompeted, and typically go extinct. Such mutations are unlikely to have a significant impact on the population dynamics. Deleterious mutations are largely irrelevant here: given the mean fitness increases by a few percent in ~80 generations, a deleterious mutation will not rise to high frequency unless it occurs contemporaneously in a cell with a large beneficial mutation, and even then is unlikely to reach high frequencies[40].

Overall population dynamics

Plotting the fitness distribution of all adaptive cells over time reveals that massive clonal interference underlies the population dynamics (Fig. 3b and Extended Data Fig. 3b). Many beneficial mutations (~20,000 observed in E1, ~11,000 observed in E2) of small s ($2\% < s < 5\%$, the 'low fitness class') drive the mean fitness early ($t < 72$), but begin to be outcompeted by cells with larger s (~10%) that stem from fewer beneficial mutations (~5,000 in E1 and ~3,000 in E2). For the first ~80 generations the mean fitness trajectory in both replicates is strikingly similar (grey curtain, Fig. 3b and Extended Data Fig 3b and Supplementary Information section 6.5). However, by ~112 generations, the mean fitness is being driven by ~100 of the most beneficial mutations ($s >$ 10%). Because mutations to these higher fitness effects are rare, they display stochastic establishment times that lead to differences in the mean fitness between the two replicates at late times (Supplementary Information section 6.5). In E2, these higher fitness mutations happen to establish earlier, contributing to a quicker decline in the low fitness class, and fewer observed adaptive lineages overall. By generation ~132, we observe that the low fitness class has shrunk to a small fraction of the population. This, however, does not mean that cells in this class are inconsequential: they prevent mutations with even smaller s from establishing. Because they are so numerous early in the evolution, some cells in this class are likely to accumulate additional beneficial mutations whose expansion could enable them to eventually outcompete cells that initially acquired higher s mutations.

Fitness effects that drive the early evolutionary dynamics in this large population are a predictable consequence of the population size and spectrum of mutation rates. The range of s at the highest frequency at time t (those that are dominating the increase in mean fitness) are those that maximize $st + \log(\mu(s))$, with $\mu(s)$ being the mutation rate to s (Supplementary Information section, 11.1). That is, the most important fitness effects at a given time are determined by a balance between being sufficiently probable to have established multiple times and sufficiently fit to have grown to large cell numbers.

Adaptive lineages that accumulate an additional beneficial mutation in a cell with an existing beneficial mutation (double mutants), can impact the dynamics. However, double mutants are rare before ~168 generations because most single mutants are not yet present at high enough cell numbers to acquire a second mutation that establishes. We estimate that fewer than ~50 of the inferred values of s and τ are impacted by double mutants (~0.2% of all adaptive lineages, Supplementary Information section 4.5). Ecological changes in the environment caused by mutants can result in frequency-dependent selection and impact the evolutionary dynamics. But, over the time range used to infer fitnesses (up to ~100 generations) our observations are consistent with the simplifying assumption that beneficial mutations have frequency-independent fitness effects and thus subpopulations only interact via competition against the mean fitness (Fig. 2a and Extended Data Fig. 2a, insets).

Discussion

Tracking a large number of small lineages provides a granular view of evolutionary dynamics that is not possible by other methods[1–3]. By focusing on sequencing just 0.002% of the genome, we gain almost five orders of magnitude in frequency resolution over genome sequencing approaches. This enables us to identify tens of thousands of independent beneficial mutations, some of which never reach frequencies above $\sim 10^{-5}$. By contrast, our previous population sequencing approach[1], which detected mutations at frequencies above ~1%, would have identified only ~15 adaptive lineages in this study (Fig. 4, Supplementary Information section 9.4). Furthermore, barcode tracking yields estimates of the fitness effects and occurrence times for all changes that convey substantial fitness advantage, whether or not they are amenable to being identified via genome sequencing.

Our results show that in an asexually evolving population of $\sim 10^8$ cells, a large number of independent beneficial mutations drive adaptation. While individually each mutation is rare and occurs stochastically, collectively they have a predictable impact on the population dynamics. In large populations therefore, the early evolutionary dynamics is almost deterministic: it only becomes stochastic when mutations so rare that they have occurred only a handful of times, or multiple mutations on the same genome, expand to an appreciable fraction of the population. Mutations with certain fitness effects play a far more important role in driving the dynamics than others, resulting in a subtle interplay between deterministic and stochastic effects.

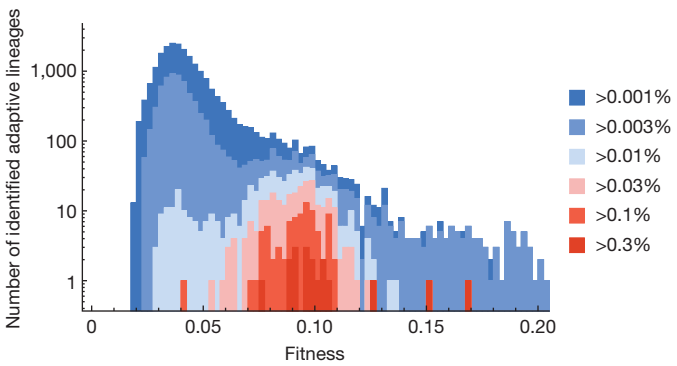


Figure 4 | The need for high frequency resolution. The fitness spectrum of adaptive lineages in replicate E1 that could be identified within the first 100 generations at different frequency resolution thresholds.

High-resolution lineage tracking is a powerful tool to study many questions important to evolution. Using this system across many environmental regimes, perhaps for longer periods of time than in this work, the relationships between adaptation rate, environment, and ecology could be quantitatively studied. A potential limitation of lineage tracking is that barcode diversity will always diminish over time. However, the possibility of adding barcodes at different times over the course of an evolution could provide a means to overcome this.

Cancer and microbial infections can have population sizes up to $\sim 10^{12}$ cells in a single individual, suggesting that massive clonal interference and complex population dynamics are likely to characterize disease progression and drug resistance[41–44]. Although mutations that rise to high frequencies are often emphasized, much larger numbers of low frequency mutations could be at least as important for disease progression or drug resistance. To study these low-frequency mutations, barcode tracking could be implemented in pathogenic microbes, cancer cell lines, or even animal tumour models[45–48]. Indeed, lineage tracking has the potential to identify the treatment regimes that most effectively slow the rate of adaptation. By randomly picking clones and sequencing their barcodes, one can cheaply identify many clones belonging to independent adaptive lineages. By sequencing the genomes of these clones, the mutational determinants for a broad range of beneficial fitness effects can be discovered. In combination with whole genome sequencing, lineage tracking therefore offers a powerful method by which to characterize the mutational spectrum underlying evolution, disease progression and drug resistance.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 11 September 2014; accepted 2 February 2015.
Published online 25 February; corrected online 11 March 2015.

1. Kvitek, D. J. & Sherlock, G. Whole genome, whole population sequencing reveals that loss of signaling networks is the major adaptive strategy in a constant environment. *PLoS Genet.* **9,** e1003972 (2013).
2. Herron, M. D. & Doebeli, M. Parallel evolutionary dynamics of adaptive diversification in *Escherichia coli*. *PLoS Biol.* **11,** e1001490 (2013).
3. Lang, G. I. *et al.* Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. *Nature* **500,** 571–574 (2013).
4. Lang, G. I., Botstein, D. & Desai, M. M. Genetic variation and the fate of beneficial mutations in asexual populations. *Genetics* **188,** 647–661 (2011).
5. Ding, L. *et al.* Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* **464,** 999–1005 (2010).
6. Shah, S. P. *et al.* Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* **461,** 809–813 (2009).
7. Mardis, E. R. *et al.* Recurring mutations found by sequencing an acute myeloid leukemia genome. *N. Engl. J. Med.* **361,** 1058–1066 (2009).
8. International Cancer Genome Consortium *et al.* International network of cancer genome projects. *Nature* **464,** 993–998 (2010).
9. Pleasance, E. D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463,** 191–196 (2010).
10. Weinreich, D. M., Delaney, N. F., DePristo, M. A. & Hartl, D. L. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* **312,** 111–114 (2006).
11. Young, B. C. *et al.* Evolutionary dynamics of *Staphylococcus aureus* during progression from carriage to disease. *Proc. Natl Acad. Sci. USA* **109,** 4550–4555 (2012).
12. Holden, M. T. G. *et al.* A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant *Staphylococcus aureus* pandemic. *Genome Res.* **23,** 653–664 (2013).
13. Desai, M. M., Walczak, A. M. & Fisher, D. S. Genetic diversity and the structure of genealogies in rapidly adapting populations. *Genetics* **193,** 565–585 (2013).
14. Neher, R. A. & Hallatschek, O. Genealogies of rapidly adapting populations. *Proc. Natl Acad. Sci. USA* **110,** 437–442 (2013).
15. Hegreness, M., Shoresh, N., Hartl, D. & Kishony, R. An equivalence principle for the incorporation of favorable mutations in asexual populations. *Science* **311,** 1615–1617 (2006).
16. Kao, K. C. & Sherlock, G. Molecular characterization of clonal interference during adaptive evolution in asexual populations of *Saccharomyces cerevisiae*. *Nature Genet.* **40,** 1499–1504 (2008).
17. Imhof, M. & Schlötterer, C. Fitness effects of advantageous mutations in evolving *Escherichia coli* populations. *Proc. Natl Acad. Sci. USA* **98,** 1113–1117 (2001).
18. Gerrits, A. *et al.* Cellular barcoding tool for clonal analysis in the hematopoietic system. *Blood* **115,** 2610–2618 (2010).
19. Desai, M. M. & Fisher, D. S. Beneficial mutation selection balance and the effect of linkage on positive selection. *Genetics* **176,** 1759–1798 (2007).

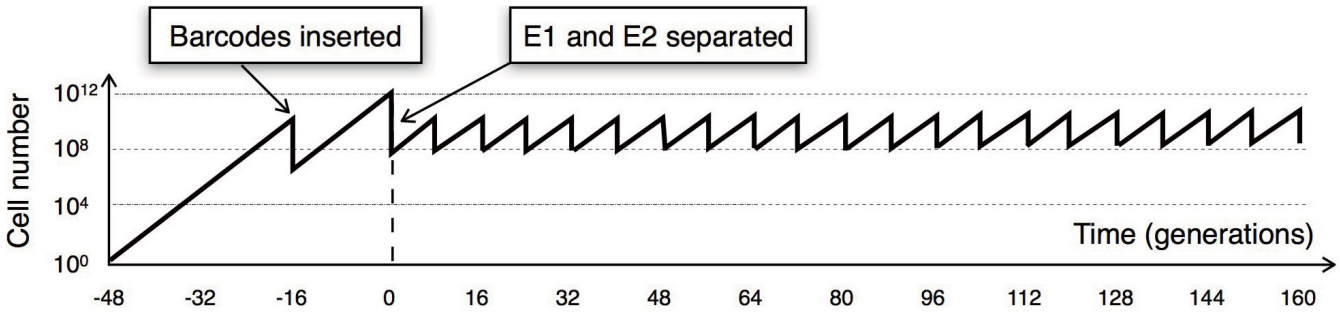
20. Charlesworth, B. The good fairy godmother of evolutionary genetics. *Curr. Biol.* **6,** 220 (1996).
21. Berns, K. *et al.* A large-scale RNAi screen in human cells identifies new components of the p53 pathway. *Nature* **428,** 431–437 (2004).
22. Smith, A. M. *et al.* Quantitative phenotyping via deep barcode sequencing. *Genome Res.* **19,** 1836–1842 (2009).
23. Lu, R., Neff, N. F., Quake, S. R. & Weissman, I. L. Tracking single hematopoietic stem cells *in vivo* using high-throughput sequencing in conjunction with viral genetic barcoding. *Nature Biotechnol.* **29,** 928–933 (2011).
24. Blundell, J. R. & Levy, S. F. Beyond genome sequencing: lineage tracking with barcodes to study the dynamics of evolution, infection, and cancer. *Genomics* **104,** 417–430 (2014).
25. Sternberg, N. & Hamilton, D. Bacteriophage P1 site-specific recombination. *J. Mol. Biol.* **150,** 467–486 (1981).
26. Austin, S., Ziese, M. & Sternberg, N. A novel role for site-specific recombination in maintenance of bacterial replicons. *Cell* **25,** 729–736 (1981).
27. Gerrish, P. J. & Lenski, R. E. The fate of competing beneficial mutations in an asexual population. *Genetica* **102–103,** 127–144 (1998).
28. Luria, S. E. & Delbrück, M. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* **28,** 491–511 (1943).
29. Lang, G. I. & Murray, A. W. Estimating the per-base-pair mutation rate in the yeast *Saccharomyces cerevisiae*. *Genetics* **178,** 67–82 (2008).
30. Lynch, M. *et al.* A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc. Natl Acad. Sci. USA* **105,** 9272–9277 (2008).
31. Zhu, Y. O., Siegal, M. L., Hall, D. W. & Petrov, D. A. Precise estimates of mutation rate and spectrum in yeast. *Proc. Natl Acad. Sci. USA* **111,** E2310–E2318 (2014).
32. Joseph, S. B. & Hall, D. W. Spontaneous mutations in diploid *Saccharomyces cerevisiae*: more beneficial than expected. *Genetics* **168,** 1817–1825 (2004).
33. Desai, M. M., Fisher, D. S. & Murray, A. W. The speed of evolution and maintenance of variation in asexual populations. *Curr. Biol.* **17,** 385–394 (2007).
34. Gillespie, J. H. Molecular evolution over the mutational landscape. *Evolution* **38,** 1116–1129 (1984).
35. Orr, H. A. The distribution of fitness effects among beneficial mutations. *Genetics* **163,** 1519–1526 (2003).
36. Kassen, R. & Bataillon, T. Distribution of fitness effects among beneficial mutations before selection in experimental populations of bacteria. *Nature Genet.* **38,** 484–488 (2006).
37. Rokyta, D. R., Joyce, P., Caudle, S. B. & Wichman, H. A. An empirical test of the mutational landscape model of adaptation using a single-stranded DNA virus. *Nature Genet.* **37,** 441–444 (2005).
38. Rokyta, D. R. *et al.* Beneficial fitness effects are not exponential for two viruses. *J. Mol. Evol.* **67,** 368–376 (2008).
39. Gresham, D. *et al.* The repertoire and dynamics of evolutionary adaptations to controlled nutrient-limited environments in yeast. *PLoS Genet.* **4,** e1000303 (2008).
40. Good, B. H., Rouzine, I. M., Balick, D. J., Hallatschek, O. & Desai, M. M. Distribution of fixed beneficial mutations and the rate of adaptation in asexual populations. *Proc. Natl Acad. Sci. USA* **109,** 4950–4955 (2012).
41. Salmon, S. E. & Smith, B. A. Immunoglobulin synthesis and total body tumor cell number in IgG multiple myeloma. *J. Clin. Invest.* **49,** 1114–1121 (1970).
42. Michaelson, J. S. *et al.* Predicting the survival of patients with breast carcinoma using tumor size. *Cancer* **95,** 713–723 (2002).
43. König, C., Simmen, H. P. & Blaser, J. Bacterial concentrations in pus and infected peritoneal fluid–implications for bactericidal activity of antibiotics. *J. Antimicrob. Chemother.* **42,** 227–232 (1998).
44. Wilson, M. L. & Gaido, L. Laboratory diagnosis of urinary tract infections in adult patients. *Clin. Infect. Dis.* **38,** 1150–1158 (2004).
45. Thomas, C. E., Ehrhardt, A. & Kay, M. A. Progress and problems with the use of viral vectors for gene therapy. *Nature Rev. Genet.* **4,** 346–358 (2003).
46. Bushman, F. *et al.* Genome-wide analysis of retroviral DNA integration. *Nature Rev. Microbiol.* **3,** 848–858 (2005).
47. Ran, F. A. *et al.* Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell* **154,** 1380–1389 (2013).
48. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339,** 823–826 (2013).
49. Kivioja, T. *et al.* Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods* **9,** 72–74 (2011).
50. Lundberg, D. S., Yourstone, S., Mieczkowski, P., Jones, C. D. & Dangl, J. L. Practical innovations for high-throughput amplicon sequencing. *Nature Methods* **10,** 999–1002 (2013).

Supplementary Information is available in the online version of the paper.

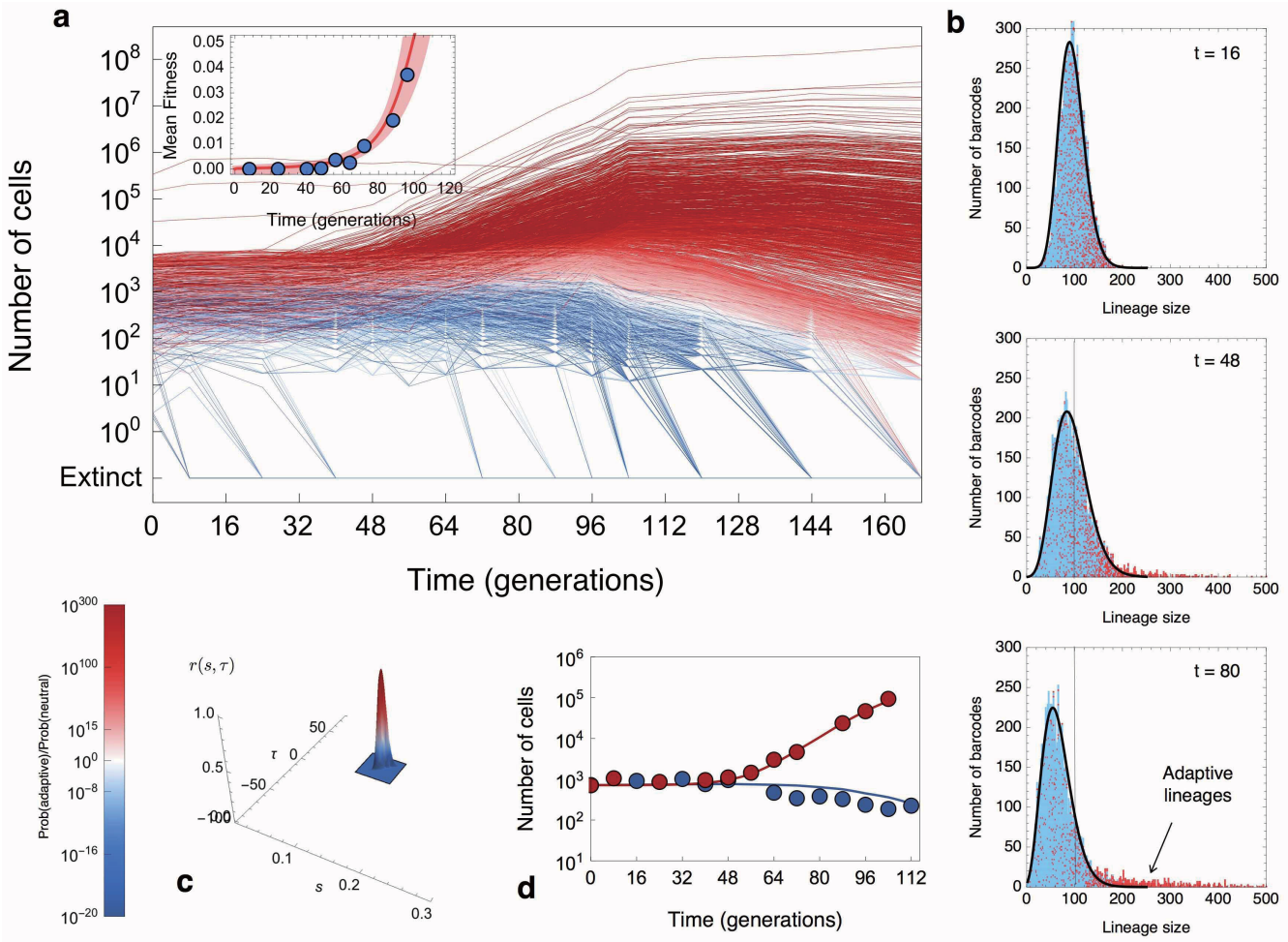
Acknowledgements The authors thank M. Siegal, K. Schwartz, B. Dunn, M. Jaffe, D. Kvitek, J. Thompson, D. Sellis, and Y. Zhu for discussions. FACS was performed at the Stanford Shared FACS Facility. We would like to acknowledge funding support from NIH grants R01 HG003328, 5-T32-HG-44-17 and R25 GM067110, NSF grants DMS-1120699 and PHY-1305433, Bio-X IIP6-63 grant from Stanford University, Gordon and Betty Moore Foundation grant no. 2919, and The Louis and Beatrice Laufer Center.

Author Contributions S.F.L. conceived of the barcoding system. S.F.L. and G.S. designed the barcoding system and evolution experiments. S.F.L., J.R.B., D.A.P., G.S. and D.S.F. developed the project vision. S.F.L. performed the barcoding and evolution experiments. S.V. and D.A.P. designed the pairwise competition assays. S.V. performed the pairwise competition assays. J.R.B. and D.S.F. developed theory and analysed the data. J.R.B. and S.F.L. wrote the paper. All authors edited the paper.

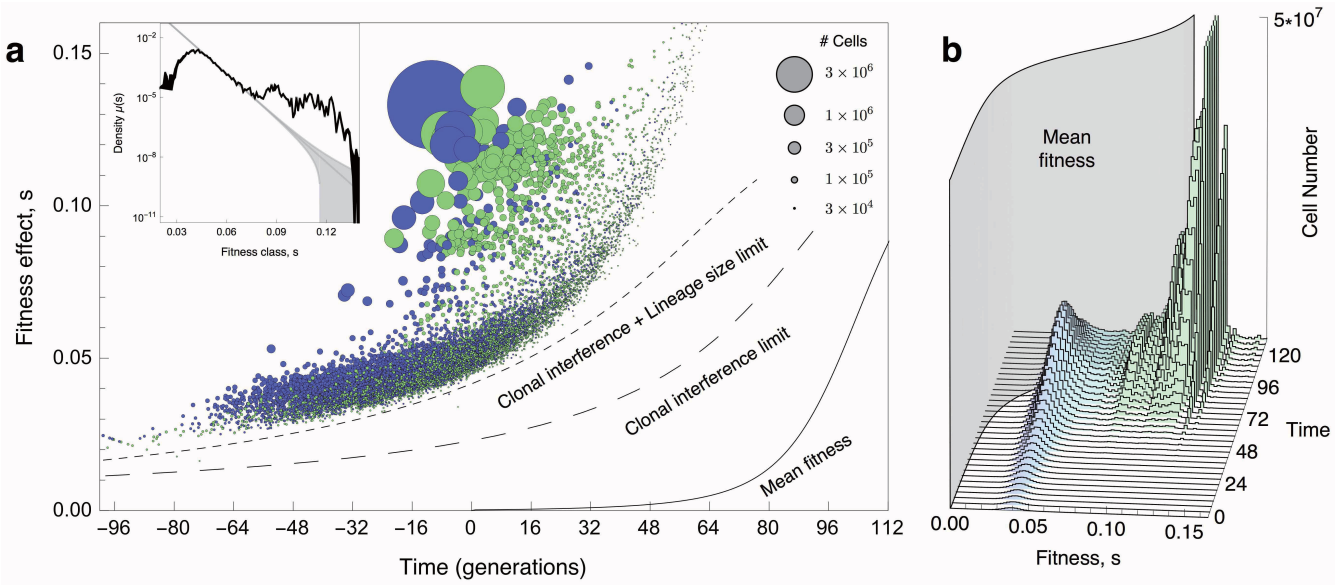
Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.S.F. (dsfisher@stanford.edu) or G.S. (gsherloc@stanford.edu).



Extended Data Figure 1 | Total population size over time. A single ancestral cell is grown for ~32 generations to ~10^{10} cells before barcodes are inserted. Cells that incorporate a barcode are grown for another 16 generations. The population is then divided into two replicates (E1 and E2) at $t = 0$. Beneficial mutations that occurred before barcoding can be sampled into both replicates.



Extended Data Figure 2 | Inferring the fitnesses and establishment times from lineage trajectories. **a**, Selected lineage trajectories and the mean fitness trajectory from replicate E2. **b**, The distribution of lineage sizes over time, for lineages that begin with ~100 ± 2 cells (vertical line). Adaptive lineages (red) begin to expand above the neutral expectation (black curve) and push neutral lineages to lower cell numbers (blue). **c**, The posterior probability distribution over s and τ for an adaptive lineage in E2. **d**, The measured trajectory of this lineage in E1 (unadaptive, blue circles) and E2 (adaptive, red circles) compared with the predicted trajectory with largest probability in E1 (blue line) and E2 (red line).



Extended Data Figure 3 | Fitness effects and establishment times for replicate E2. **a**, Scatter plot of τ and s of all ~14,000 beneficial mutations (circles) identified in E2. Circle area represents the size of the lineage at generation 88. Purple circles indicate lineages with mutations that occurred in the period of common growth ($t < 0$) that were sampled into, and established in, E1 and E2. Green circles indicate lineages that were identified as adaptive in only one replicate and likely contain mutations that arose after $t = 0$. Lines indicate the time limits before which mutations must occur in order to establish (large dash) or be observed (small dash). These limits trail the mean fitness (solid line) by $\sim 1/s$ generations. Inset, the spectrum of mutation rates, $\mu(s)$, as a function of fitness effect, **s** inferred from mutations that likely occurred after $t = 0$ (Supplementary Information section 10.2). The y axis is the mutation rate density, so the mutation rate to a range, Δs, is obtained by multiplying this by Δs. The total beneficial mutation rate to $s > 5\%$ is inferred to be $\sim 1 \times 10^{-6}$ and is consistent across replicates. The observed spectrum is not exponential (grey line, with the error range shaded). **b**, The distribution of the number of adaptive cells binned by their fitness over time. As the mean fitness (grey curtain) surpasses the fitness of a subpopulation, cells with that fitness begin to decline in frequency.

ARTICLE

doi:10.1038/nature14259

Notum deacylates Wnt proteins to suppress signalling activity

Satoshi Kakugawa[1]*, Paul F. Langton[1]*, Matthias Zebisch[2]†*, Steven A. Howell[1], Tao-Hsin Chang[2], Yan Liu[3], Ten Feizi[3], Ganka Bineva[4], Nicola O'Reilly[4], Ambrosius P. Snijders[5], E. Yvonne Jones[2] & Jean-Paul Vincent[1]

Signalling by Wnt proteins is finely balanced to ensure normal development and tissue homeostasis while avoiding diseases such as cancer. This is achieved in part by Notum, a highly conserved secreted feedback antagonist. Notum has been thought to act as a phospholipase, shedding glypicans and associated Wnt proteins from the cell surface. However, this view fails to explain specificity, as glypicans bind many extracellular ligands. Here we provide genetic evidence in *Drosophila* that Notum requires glypicans to suppress Wnt signalling, but does not cleave their glycophosphatidylinositol anchor. Structural analyses reveal glycosaminoglycan binding sites on Notum, which probably help Notum to co-localize with Wnt proteins. They also identify, at the active site of human and *Drosophila* Notum, a large hydrophobic pocket that accommodates palmitoleate. Kinetic and mass spectrometric analyses of human proteins show that Notum is a carboxylesterase that removes an essential palmitoleate moiety from Wnt proteins and thus constitutes the first known extracellular protein deacylase.

Negative feedback characterizes biological signalling[1] and although often cell-intrinsic, is also mediated by secreted proteins. Cell- and non-cell-autonomous feedbacks modulate signal transduction by Wnt proteins, a class of secreted proteins characterized by the presence of palmitoleic acid appended on a conserved serine[2,3]. This palmitoleic acid moiety is essential for signalling[2,4,5], contributing to interaction with Frizzled receptors[3,6,7]. Canonical Wnt signalling triggers expression of intracellular, extracellular and membrane-localized inhibitors of the pathway. Secreted inhibitors include Dickkopf (Dkk) family members, which bind to the extracellular domain of the Wnt co-receptor low-density-lipoprotein-receptor-related protein 5/6 (Lrp5/6), as well as Wnt inhibitory factor 1 (Wif1) and secreted Frizzled receptor proteins (Sfrp), which sequester Wnt proteins[8]. Tiki is a membrane-bound protease that cleaves the amino-terminal region of Wnt ligands[9]. Notum is also thought to act enzymatically[10,11] but on glypicans, a class of heparan sulfate proteoglycans (HSPGs) implicated in the extracellular stabilization, movement, and/or surface retention of Wnt proteins, as well as of other signalling ligands[12–14].

Notum orthologues are found in metazoans from planarians to humans and all bear the hallmark Ser-His-Asp catalytic triad of α/β-hydrolases[10,11]. The sequence similarity of Notum to plant pectin acetylesterases prompted the early suggestion that it could hydrolyse glycosaminoglycan (GAG) chains of glypicans[10,11], thus affecting their ability to interact with Wnt ligands and somehow modulating signalling activity. It was subsequently reported that Notum triggers glypican shedding from cultured cells, perhaps by cleaving their glycosylphosphatidylinositol (GPI) anchor[15,16]. Indeed, the currently accepted view is that Notum is a glypican-specific phospholipase[17]. However, glypican-based interactions also modulate Dpp (*Drosophila* TGF-β), Hedgehog, and fibroblast growth factor, as well as Wingless signalling[12–14]. One would expect therefore that these pathways would also be sensitive to Notum-induced glypican release. Yet, existing evidence suggests that Notum is primarily a feedback inhibitor of Wnt signalling. In planarian worms, *Drosophila*, zebrafish and hepatocarcinomas, *notum* expression is activated by Wnt signalling and, conversely, Notum seems to preferentially suppress Wnt signalling[10,11,18–21]. Because more pleiotropic effects would be expected from an enzyme that targets glypicans, we felt compelled to reassess Notum's specificity and mode of action.

Notum specifically inhibits Wnt signalling

To investigate the specificity of Notum systematically, we analysed its effects on *Drosophila* wing imaginal discs, which require Wingless (the main *Drosophila* Wnt), Dpp and Hedgehog for patterning and growth[22,23]. As expected, overexpression of *Drosophila* Notum (dNotum) throughout the dorsal compartment prevented expression of *senseless*, a gene normally activated by high level Wingless signalling. By contrast, *patched* (*ptc*), a Hedgehog target gene[24,25], was unaffected (Fig. 1a, b) and phospho-Mad immunoreactivity, a marker of Dpp signalling[26], was only mildly reduced (Extended Data Fig. 1a, b). Loss-of-function assays, in homozygous *notum* knockout ($notum^{KO}$) tissue, confirmed the specificity of dNotum to Wingless signalling (Fig. 1c and Extended Data Fig. 1c). Although complete loss of *notum* was lethal, strong hypomorphic animals ($notum^{141}/notum^{KO}$) survived to adulthood. The wings of such animals had supernumerary margin bristles, consistent with excess Wingless signalling, but had no defects indicative of impaired Hedgehog or Dpp signalling (Extended Data Fig. 1d–g). Nevertheless, extensive evidence suggests that glypicans contribute to these two signalling pathways[27–30]. This is difficult to reconcile with the apparent specificity of Notum if it acts as a glypican-specific phospholipase.

Notum does not cleave the GPI anchor of glypicans

One previously reported observation, namely that dNotum inhibits signalling by membrane-tethered (that is, shedding-resistant) Wingless[11] (Extended Data Fig. 2a, b), is incompatible with the view that Notum is a glypican-specific phospholipase. In addition, genetic removal of the two *Drosophila* glypicans Dally and Dally-like protein (Dlp) did not

[1]MRC's National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK. [2]Division of Structural Biology, Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK. [3]Glycosciences Laboratory, Imperial College London, Department of Medicine, Du Cane Road, London W12 0NN, UK. [4]Cancer Research UK, London Research Institute, 44 Lincoln's Inn Fields, London WC2A 3LY, UK. [5]Cancer Research UK, Clare Hall Laboratories, Blanche Lane, South Mimms, Potters Bar, Hertfordshire EN6 3LD, UK. †Present address: Evotec (UK) Ltd, 114 Innovation Drive, Milton Park, Abingdon, Oxfordshire OX14 4RZ, UK.
*These authors contributed equally to this work.

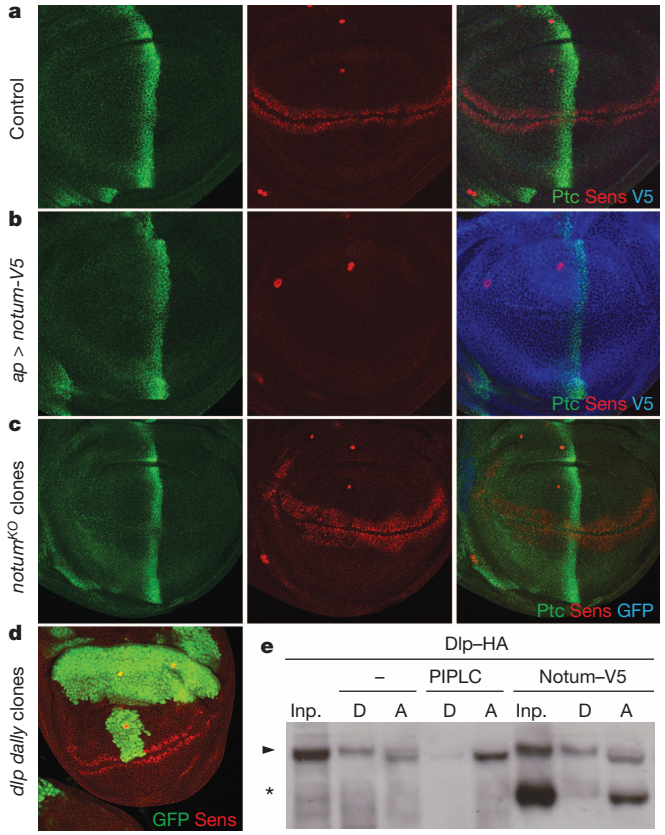


Figure 1 | Notum specifically inhibits Wnt signalling. **a**, **b**, Overexpression of V5-tagged dNotum (Notum–V5) with the *apterous-Gal4* driver, which is expressed in the dorsal compartment prevents expression of Senseless (Sens) but not that of Patched (Ptc) (**b**). **c**, Loss of *notum* activity, achieved by generating large patches of $notum^{KO}$ tissue (see Methods), marked by the loss of green fluorescent protein (GFP), leads to broadening of Senseless expression but does not affect Patched expression. As in all subsequent confocal images, third instar wing imaginal discs are shown with posterior to the right and dorsal up. **d**, Senseless is expressed seemingly normally in large patches of *dlp dally* mutant cells (GFP-negative). **e**, Western blot (co-stained with anti-V5 and anti-haemagglutinin (HA)) of phase-separated extracts of S2 cells transfected with a plasmid expressing HA-tagged Dlp (Dlp–HA). In control extracts, Dlp (arrowhead) is found equally in the detergent (D) and aqueous (A) phases. Coexpression of dNotum–V5 (asterisk) had no impact while treatment with PIPLC shifted Dlp to the aqueous phase.

abrogate high level Wingless signalling (Fig. 1d). These two sets of data strongly suggest that glypican shedding is unlikely to account for the inhibitory effect of dNotum on Wingless signalling. Indeed, we could not reproduce the results of an earlier phase partition assay, which suggested that Notum increases the water solubility of glypicans, as expected from GPI cleavage[15]. Extracts from cells expressing tagged Dlp or Dally were treated with either dNotum or bacterial phosphoinositide phospholipase C (PIPLC), an enzyme known to cleave GPI anchors. PIPLC caused both glypicans to partition almost exclusively in the aqueous phase, but dNotum did not (Fig. 1e for Dlp; Extended Data Fig. 2c for Dally), even though, under these conditions, it was effective at inhibiting signalling (Extended Data Fig. 2d). Likewise, in imaginal discs, Notum did not mimic PIPLC: whereas extracellular Dlp and Dally were noticeably reduced after addition of exogenous PIPLC, overexpression of dNotum had no such effect (Extended Data Fig. 2e–l). Therefore, experiments with cultured cells and imaginal discs suggest that Notum is not a glypican-specific phospholipase.

Glypicans contribute to the activity of Notum

Although dNotum does not seem to modulate Wingless signalling by cleaving the GPI anchor of glypicans, genetic interactions between *notum* and *dlp* suggest a functional relationship[31,32]. We therefore investigated the role of Dlp or Dally in the ability of dNotum to suppress Wingless signalling. dNotum overexpression, along the anterior–posterior (A–P) boundary, led to complete and long-range suppression of Senseless expression (Fig. 2a). In the absence of either Dlp or Dally, this activity was very much reduced, as indicated by the recovery of endogenous Senseless expression (Fig. 2b, c). Notably, Dally was also required for Notum to suppress signalling by membrane-tethered Wingless (Extended Data Fig. 3a). Because Dally is not essential for survival, its requirement for Notum's ability to suppress Wingless signaling could be confirmed in adult wings (Extended Data Fig. 3b–d). To address the relevance of glypican GPI anchorage, we created a transgene expressing Dlp-CD8 (34 carboxy-terminal amino acids of Dlp replaced by the CD8 transmembrane domain) under control of the *tubulin* promoter. This transgene restored the ability of overexpressed dNotum to repress Wingless signalling in *dlp* mutant homozygotes (Fig. 2d; compare to Fig. 2b), confirming the importance of glypicans but not their GPI anchor.

Glypicans bear sulfated glycans. In *Drosophila*, sulfation of the sugar chains requires Sulfateless, a GlcNAc *N*-deacetylase/*N*-sulfotransferase (NDST)[33], which can be knocked down *in vivo* with an RNA interference (RNAi) transgene. Gal4 was used to express this transgene specifically in the posterior compartment, leaving the anterior compartment as a control. At the same time, a *dpp-LexA* driver was used to overexpress dNotum along the A–P boundary. Overexpressed dNotum inhibited Senseless expression in the control compartment but not in the territory deficient in *sulfateless* activity (Fig. 2e). Therefore, sulfation of HSPGs is needed for dNotum to act. Notably, overexpressed dNotum did not accumulate in the compartment expressing the *sulfateless* RNAi transgene (compare Fig. 2e to Fig. 2a, right panels). Likewise, dNotum was depleted from the surface of *dally dlp* double-mutant cells generated by mitotic recombination (Fig. 2f). These findings suggest that Dally and Dlp retain dNotum at the cell surface through interaction with their sulfated glycans. Indeed, dNotum bound specifically to sulfated glycans on a glycan array (Extended Data Fig. 4). In addition, surface plasmon resonance (SPR) showed that recombinant human (h) $NOTUM_{core}$ (Ser 81–Thr 451, Cys330Ser) bound to heparin and heparan sulfate with micromolar affinities. The dissociation constant of a complex comprising $hNOTUM_{core}$ and human glypican-3 (GPC3 Pro 31–Asn 538) was 104 μM (Fig. 3a). Consistent with the *Drosophila* genetic data, this binding relies largely on the two sulfated GAG chains in GPC3 as their removal led to a more than fivefold reduction in affinity (Fig. 3a). We conclude that sulfated GAG chains on glypicans probably mediate their interaction with Notum.

Structure-guided identification of GAG-binding sites

The above results indicate that glypicans contribute to Notum activity by localizing it at the cell surface, but are unlikely to be the target of Notum's enzymatic activity. What could the target be? We started to address this question by solving the structures of $hNOTUM_{core}$ (in nine crystal forms at resolutions between 1.4 and 2.8 Å: see Supplementary Information) and of $dNotum_{\Delta loop}$ (in two crystal forms at resolutions of 2.4 and 1.9 Å) (Fig. 3b, Extended Data Fig. 5 and Supplementary Information). The structures exhibit a canonical α/β-hydrolase fold[34], as predicted[10,11]. The conserved eight-stranded central β-sheet is extended on both sides by strands 4 and 14 and is flanked by the canonical six α-helices. This single domain topology is further extended by additional α-helices, two very short β-sheets, several long loops and seven stabilizing disulfides. The catalytic triad comprises Ser 232, Asp 340 and His 389 (hNOTUM residue numbering).

Seven sulfate binding sites were identified in $hNOTUM_{core}$ crystal form III (Fig. 3c and Extended Data Fig. 6). Among them, one (sulfate 1) was found by SPR to contribute substantially to heparin–Notum interactions (Fig. 3d). In addition, co-crystals with short heparin oligomers or sucrose octasulfate (SOS), a heparin mimic, were generated and analysed. These structural studies and additional biophysical analyses (described in Supplementary Information and illustrated in Extended Data

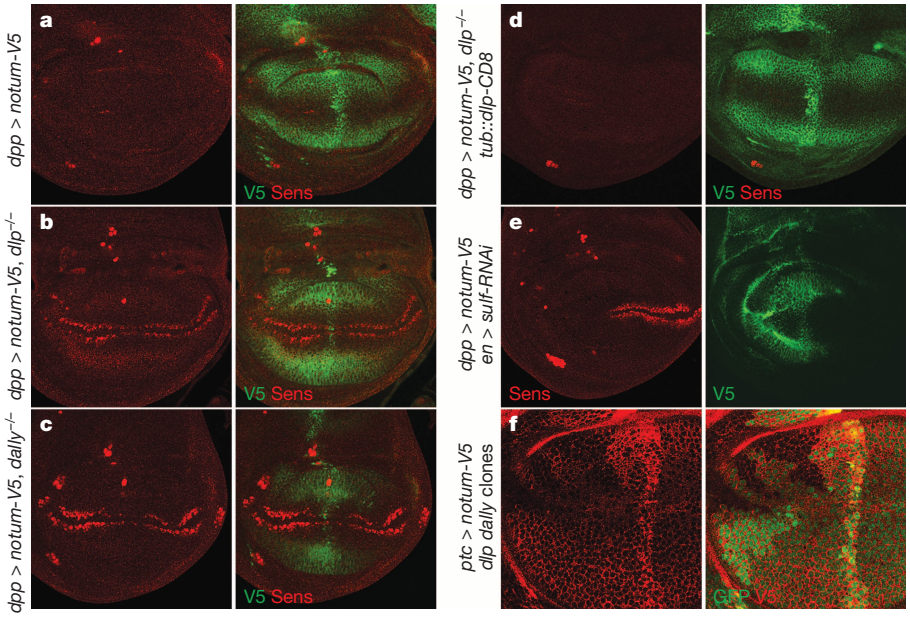


Figure 2 | Notum requires the GAGs of glypicans to inhibit Wingless signalling. **a**–**c**, Ectopically expressed dNotum–V5 does not suppress Senseless expression in the absence of *dlp* (**b**) or *dally* (**c**). Although dNotum is only expressed in a vertical band along the A–P boundary, it spreads along the whole A–P axis. **d**, Ectopic dNotum represses Senseless expression in *dlp* mutants that express Dlp-CD8 (*tubulin* promoter). **e**, Expression of an RNAi transgene against *sulfateless* (*sulf-RNAi*) in the posterior compartment prevents dNotum–V5 (expressed from *dpp-lexA lex-op-notum-V5*) from being retained at the cell surface and from suppressing Senseless expression. Wingless signalling is still suppressed in the anterior compartment. **f**, Accumulation of dNotum–V5 is reduced at the surface of *dlp dally* double-mutant tissue (GFP-negative).

Fig. 6) delineated an extensive GAG-binding patch centred on a basic groove between the top of the β-sheet and helix αK (Fig. 3c). Importantly, the GAG-binding surface on Notum is distant from the catalytic triad, consistent with our earlier evidence that Notum binds to glypicans, but does not act on them enzymatically.

Evidence for carboxylesterase activity

The α/β-hydrolase superfamily includes proteases, lipases, esterases, dehalogenases, peroxidases and epoxide hydrolases[34]. To identify which of these activities relate most closely to the activity of the Notum protein, we compared the structure of hNOTUM to those of all known α/β-hydrolases (PDBeFold server[35]). The search returned many weak homologues, including human esterase D[36] and acyl-protein thioesterase 1 (APT1)[37] (Extended Data Fig. 7a). A structure-based search for function using the ProFunc Server[38] also suggested that Notum is a carboxylesterase. Furthermore, the closest non-animal homologues of Notum, the pectin acetylesterases of angiosperms (22% sequence identity to hNOTUM, Extended Data Fig. 7b) are carboxylesterases. We assessed the functional significance of these observations by measuring the activity of $hNOTUM_{core}$ on *p*-nitrophenyl (pNP) acetate (pNP2), a chromogenic carboxylesterase substrate[39]. Pronounced activity could be detected (Fig. 4a). This activity was strongly inhibited by Triton X-100, and by phenylmethanesulfonyl fluoride (PMSF), a compound known to covalently modify the catalytic serine of serine esterases and proteases (Extended Data Fig. 8a, b). By contrast, there was no measurable hNOTUM activity on representative sulfatase, phosphatase, phospholipase C or amidase/protease substrates (Fig. 4a). Addition of SOS or heparin resulted in a modest increase in Notum carboxylesterase activity (Extended Data Fig. 8a). The possibility that GAGs also contribute to Notum function by allosteric activation requires further investigation.

As a secreted carboxylesterase that inhibits Wnt signalling, Notum is likely to target a carboxy-oxoester or carboxy-thioester bond present on an extracellular component of the Wnt signal transduction machinery. The linkage between Wnt and palmitoleic acid is the only such

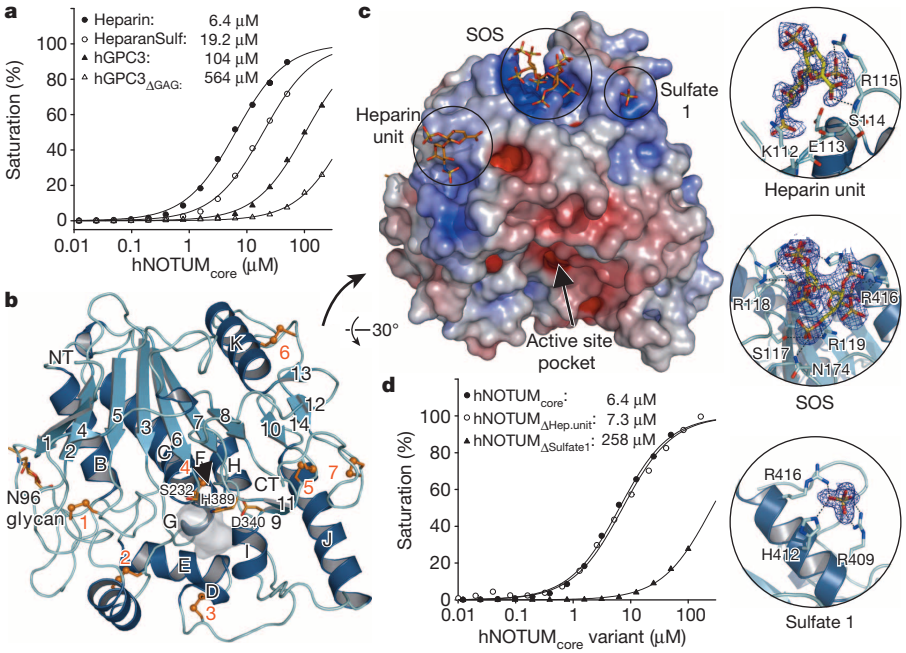


Figure 3 | hNOTUM structure and GAG binding. **a**, Binding of $hNOTUM_{core}$ to immobilized heparin, heparan sulfate (HeparanSulf), hGPC3 or $hGPC3_{\Delta GAG}$, assayed by SPR. **b**, Structure of hNOTUM. β-strands are numbered and α-helices are labelled alphabetically from N to C terminus (NT and CT, respectively). Disulfides are shown in orange, catalytic triad residues as sticks and the active site pocket shaded grey. Asn 96 is glycosylated (also in dNotum). **c**, Heparin-mimicking ligands from three different structures are plotted onto a surface representation coloured by electrostatic potential from red ($-8k_bT/e_c$) to blue ($-8k_bT/e_c$). Close-up views of binding sites are shown on the right with experimental omit electron density contoured at 2.0σ. **d**, SPR assay measuring $hNOTUM_{core}$ variant binding to immobilized heparin. Mutation of the heparin disaccharide binding site (Arg115Ser; $hNOTUM_{\Delta Hep.unit}$) had little effect while mutations in the sulfate binding site 1 (Arg409Gln, His412Asn and Arg416Gln; $hNOTUM_{\Delta Sulfate1}$) strongly reduced binding. For SPR (**a**, **d**), each data point is the mean result of two replicates.

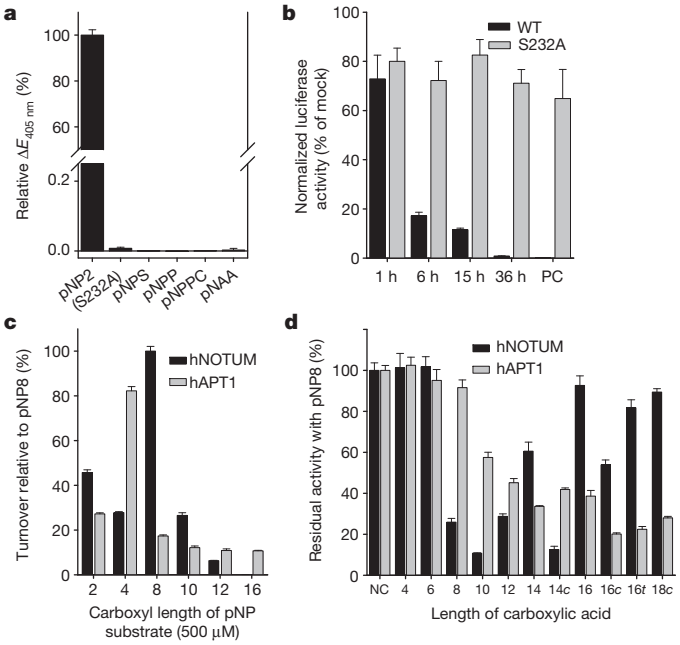


Figure 4 | Enzymatic activity of hNOTUM. **a**, Activity of $hNOTUM_{core}$ and its Ser232Ala variant on *p*-nitrophenyl (pNP) acetate (pNP2) and activity of $hNOTUM_{core}$ on other chromogenic substrates. pNAA, *p*-nitroacetanilide (amidase/protease substrate); pNPP, pNP-phosphate (phosphatase substrate); pNPPC, pNP-phosphorylcholine (phospholipase C substrate); pNPS, pNP-sulfate (sulfatase substrate). **b**, mWnt3A inactivation by hNOTUM. After the indicated time (in hours), $hNOTUM_{core}$ or its Ser232Ala variant was removed with cobalt affinity beads and residual Wnt3A activity measured with TOPFlash. PC denotes no hNOTUM removal. Results are normalized to those from identically treated mock samples. **c**, Activity of hNOTUM and hAPT1 on chromogenic *p*-nitrophenyl ester substrates of different lengths. **d**, Inhibition of hNOTUM by various carboxylic acids. pNP8 was used as substrate at a concentration of 1 mM, as were the carboxylic acids. *c* or *t* denote *cis* or *trans* C9–C10 double bond. All graphs show the mean ± s.d. ($n = 4$).

chemical bond described to date, suggesting that Notum could target Wnt proteins themselves. To evaluate this possibility, we treated mouse (m)Wnt3A with recombinant $hNOTUM_{core}$ for specific durations, removed the $hNOTUM_{core}$ and used a cell-based luciferase assay to measure signalling activity of the remaining mWnt3A. This showed that hNOTUM inactivated mWnt3A directly, irreversibly and in a time-dependent manner (Fig. 4b), while no such effect could be detected on Norrin, a non-lipidated ligand that also acts via the Wnt receptors[40] (Extended Data Fig. 8c).

Remarkably, the Notum crystal structures revealed a large (~380 $Å^3$), hydrophobic pocket adjacent to the catalytic triad (Fig. 3b, c). Computational docking showed that this pocket could accommodate long-chain fatty acids of up to 16 carbon atoms (C16). The size restriction imposed on saturated fatty acids was functionally assessed by measuring hNOTUM enzymatic activity on commercially available saturated chromogenic pNP ester substrates of varying chain lengths. The activity of human APT1, a cytosolic thio- and oxoesterase was measured in parallel for comparison. hNOTUM had a pronounced preference for pNP8 (Fig. 4c), with a micromolar Michaelis constant (Extended Data Fig. 8d, e). The activity for pNP-palmitate (pNP16) was less than 0.2% of that for pNP8. To extend our studies of hNOTUM specificity beyond commercially available substrates, we used a competitive inhibition assay, using pNP8 as substrate. Saturated 8–12 carbon (C8-C12) long linear carboxylic acids inhibited activity (Fig. 4d) while longer saturated fatty acids had no effect. Interestingly, however, strong inhibition was observed with the Wnt-associated *cis*-unsaturated lipids myristoleic (C14) and palmitoleic acid (C16) (Fig. 4d and Extended Data Fig. 8f), but not with palmitelaidic acid, the *trans* isomer of the 16:1 fatty acid (Fig. 4d). These results confirm that Notum can bind to C14 and C16 carboxylic acids if they contain a C9–C10 *cis* double bond and therefore might hydrolyse the oxo-ester bond linking palmitoleate or myristoleate to Wnt proteins.

Notum deacylates Wnt proteins

To test directly Notum-mediated Wnt deacylation we turned to liquid chromatography–mass spectrometry (LC–MS) analysis. mWnt3A was purified from conditioned medium, treated with recombinant $hNOTUM_{core}$ or a mock solution, differentially isotope labelled, and trypsinised. No notable identification could be obtained for the predicted palmitoleoylated tryptic peptide, indicating incompatibility with the LC–MS conditions. After treatment with hNOTUM, however, this peptide could be identified and quantified in non-acylated form (Fig. 5a, b and Extended Data Fig. 9a, b). Replicate LC–MS measurements and label reversal consistently showed an increase in signal intensity for the hNOTUM-treated de-acylated peptide whereas control peptides were largely unaffected by hNOTUM treatment (Extended Data Fig. 9c, d). This suggests that treatment of mWnt3A with hNOTUM removes the palmitoleic acid moiety thus rendering the relevant peptide more hydrophilic and detectable by LC–MS. Encouraged by these results, we proceeded to assess the activity of hNOTUM on synthetic peptides. The predicted tryptic peptide from hWNT3A was synthesized in a disulfide-bonded form with a palmitoleate group on the relevant serine (Supplementary Information). These peptides were treated with recombinant $hNOTUM_{core}$, or with $hNOTUM_{core}$(S232A), which is predicted to be enzymatically inactive, and the reaction products were analysed by matrix assisted laser desorption ionization time-of-flight (MALDI-TOF). No significant deacylation was detected in $hNOTUM_{core}$(S232A)-treated samples, whereas hNOTUM-treated peptides were found to be extensively deacylated (Fig. 5c and Extended Data Fig. 9e). We conclude from these assays that Notum catalyses the removal of palmitoleic acid, which is normally *O*-linked to Ser 209 of hWNT3A. We also assayed the effect of hNOTUM on a synthetic peptide from human Sonic Hedgehog (SHH), which is *N*-palmitoylated at the amino terminus[41]. No change in the level of acylation could be detected (Fig. 5d and Extended Data Fig. 9f), confirming that the activity of Notum on Wnt is specific, in agreement with our genetic evidence.

To gain structural insight into Wnt–Notum recognition, we co-crystallized inactive $hNOTUM_{core}$(S232A) with a palmitoleoylated disulfide-bonded peptide corresponding to hWNT7A(Cys 202–Cys 209). The crystal structure revealed the palmitoleoyl group occupying the active site pocket (Fig. 5e and Extended Data Fig. 9g). Electron density was also evident for the ester bond. No interpretable density was found for the peptide, probably owing to disorder. This apparent lack of interaction with the peptide concurs with the general observation that esterases/lipases of the α/β-hydrolase family bind only to the acid part of the ester substrate. The carboxylic acid carbon is 3.3 Å from the Cβ of the mutated serine nucleophile, a distance consistent with ideal positioning of the hydroxyl for nucleophilic attack. Classically esterase-catalysed hydrolysis proceeds through a tetrahedral transition state characterized by a negatively charged carbonyl oxygen stabilized by two canonical backbone amides, the oxyanion hole[34]. In hNOTUM, the Gly 127–Trp 128 amide participates in formation of the oxyanion hole in addition to the canonical Ser 232–Ala 233 and Gly 126–Gly 127 amides, thereby providing optimal stabilization during the transition state (Extended Data Fig. 9g). The kinked *cis* double bond (C9–C10) of the acyl tail is positioned at the base of the pocket between Ile 291, Phe 319 and Phe 320. We found a similar binding mode for a hNOTUM–myristoleate crystal structure (Extended Data Fig. 9h). Thus, the binding pocket can accommodate extended carbon tails up to C8/C10 but longer fatty acid chains must be kinked at this point in order to fit in. Saturated fatty acids generally adopt an extended conformation, explaining the preference of Notum for palmitoleate and myristoleate (both *cis*-unsaturated lipids kinked at C9–C10) over palmitate and myristate. The pocket entrance (lined by Ser 232 and His 389) is relatively narrow, but

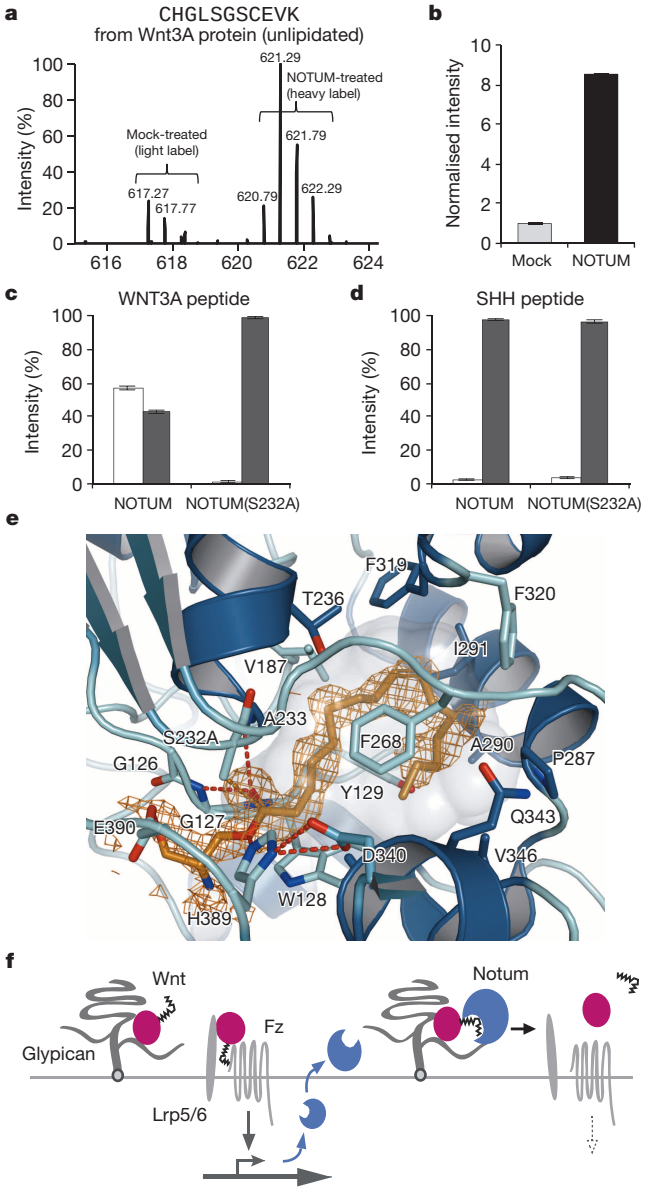


Figure 5 | Wnt-deacylation by Notum. **a**, LC–MS analysis of mWnt3A protein treated with $hNOTUM_{core}$ or a mock solution. By comparison to mock treatment (light label), addition of hNOTUM (heavy label) caused a significant increase in the signal intensity of unlipidated CHGLSGSCEVK. **b**, LC–MS peak areas from **a**, shown as mean ± s.e.m. ($n = 2$). **c**, **d**, Quantification from MALDI analysis of synthetic lipid-bearing peptides treated with $hNOTUM_{core}$ or its Ser232Ala variant. Bars (grey denotes lipidated; white denotes delipidated) show mean ± s.e.m. ($n = 3$). Palmitoleoylated hWNT3A peptide, but not palmitoylated hSHH peptide, was specifically deacylated by the wild-type enzyme. **e**, Close-up view on the seryl-palmitoleate active site complex of hNOTUM. The experimental omit electron density is contoured at 2σ. **f**, Feedback control by Notum. Notum deacylates Wnt in a glypican-assisted fashion.

comparisons of all hNOTUM structures suggest substantial flexibility, compatible with palmitoleate entry and release (Extended Data Fig. 5b). Therefore, crystallographic evidence strengthens our observation that Notum is a Wnt-specific deacylase with preference for *cis*-unsaturated long chain lipids.

Discussion

Only a small number of secreted proteins, Wnts, Hedgehogs and Ghrelins, are known to be acylated[42]. In all cases, this post-translational modification is essential for activity and is carried out by dedicated membrane-bound *O*-acyl transferases (MBOATs). Porcupine, the Wnt MBOAT, appends palmitoleate and shorter *cis*-unsaturated fatty acids onto Wnt[43]. We have shown here that Notum specifically deacylates Wnt (Fig. 5f) and is thus the first enzyme known to deacylate an extracellular protein. The specificity of Notum can be traced to the shape of its hydrophobic pocket, which can accommodate *cis*-unsaturated fatty acids such as myristoleate and palmitoleate, and the nature of its enzymatic activity, a carboxyl oxoesterase. These characteristics ensure that Notum preferentially acts on Wnt proteins, the only secreted proteins known to be *O*-palmitoleoylated on a serine residue. Notum enzymatically inhibits signalling activity by removing the palmitoleate moiety of Wnt proteins, which contributes directly to receptor binding[3]. Notum could also interfere non-catalytically with the formation of the Wnt–Frizzled complex by sequestering the palmitoleate moiety as overexpressed dNotum(S237A) mildly suppressed Wingless signalling *in vivo* (data not shown). We have found that glypicans are required for Notum function and that Notum binds to the sulfated GAGs of HSPGs. Glypicans can have stimulatory roles in Wnt signalling[44,45]. However, in the presence of Notum, we suggest that glypicans are also inhibitory by acting as a scaffold that co-localizes Notum and its substrate (Wnts) at the cell surface (Fig. 5f).

Our results point to Notum's physiological targets being exclusively Wnt family members. Notum is the only secreted Wnt feedback inhibitor found across the metazoan kingdom, from planarians to humans, although it is seemingly absent from *Caenorhabditis elegans*. Notum's Wnt-deacylation activity, along with other means of feedback inhibition such as ligand sequestration, receptor blocking, receptor down-regulation and proteolytic degradation[9,46–48] undoubtedly contributes to the fine balancing of Wnt signalling both during development, for cell fate specification, and in adults, for example, for stem cell maintenance. Indeed, insufficient or excessive Wnt signalling has been associated with diseases such as neurodegeneration or cancer, respectively. Our binding data suggest that Notum could possibly be modulated by dietary *cis*-unsaturated fatty acids. Moreover, because Notum is an extracellular enzyme with a well-defined and large active site pocket, it is probably amenable to chemical inhibition to alleviate conditions associated with insufficient Wnt signalling. Conversely, recombinant Notum could be considered as a therapeutic agent to prevent excess Wnt signalling such as in Wnt-driven cancers.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 21 October 2014; accepted 26 January 2015.
Published online 25 February 2015.

1. Freeman, M. Feedback control of intercellular signalling in development. *Nature* **408,** 313–319 (2000).
2. Takada, R. *et al.* Monounsaturated fatty acid modification of Wnt protein: its role in Wnt secretion. *Dev. Cell* **11,** 791–801 (2006).
3. Janda, C. Y., Waghray, D., Levin, A. M., Thomas, C. & Garcia, K. C. Structural basis of Wnt recognition by Frizzled. *Science* **337,** 59–64 (2012).
4. Willert, K. *et al.* Wnt proteins are lipid-modified and can act as stem cell growth factors. *Nature* **423,** 448–452 (2003).
5. Tang, X. *et al.* Roles of N-glycosylation and lipidation in Wg secretion and signaling. *Dev. Biol.* **364,** 32–41 (2012).
6. Clevers, H. & Nusse, R. Wnt/& β-catenin signaling and disease. *Cell* **149,** 1192–1205 (2012).
7. Kim, S. E. *et al.* Wnt Stabilization of β−Catenin Reveals Principles for Morphogen Receptor-Scaffold Assemblies. *Science* **340,** 867–870 (2013).
8. Niehrs, C. The complex world of WNT receptor signalling. *Nature Rev. Mol. Cell Biol.* **13,** 767–779 (2012).
9. Zhang, X. *et al.* Tiki1 is required for head formation via Wnt cleavage-oxidation and inactivation. *Cell* **149,** 1565–1577 (2012).
10. Giráldez, A. J., Copley, R. R. & Cohen, S. M. HSPG modification by the secreted enzyme Notum shapes the Wingless morphogen gradient. *Dev. Cell* **2,** 667–676 (2002).
11. Gerlitz, O. & Basler, K. Wingful, an extracellular feedback inhibitor of Wingless. *Genes Dev.* **16,** 1055–1059 (2002).
12. Filmus, J., Capurro, M. & Rast, J. Glypicans. *Genome Biol.* **9,** 224 (2008).
13. Yan, D. & Lin, X. Shaping morphogen gradients by proteoglycans. *Cold Spring Harb. Perspect. Biol.* **1,** a002493 (2009).

14. Bornemann, D. J., Duncan, J. E., Staatz, W., Selleck, S. & Warrior, R. Abrogation of heparan sulfate synthesis in *Drosophila* disrupts the Wingless, Hedgehog and Decapentaplegic signaling pathways. *Development* **131,** 1927–1938 (2004).
15. Kreuger, J., Perez, L., Giraldez, A. J. & Cohen, S. M. Opposing activities of Dally-like glypican at high and low levels of Wingless morphogen activity. *Dev. Cell* **7,** 503–512 (2004).
16. Traister, A., Shi, W. & Filmus, J. Mammalian Notum induces the release of glypicans and other GPI-anchored proteins from the cell surface. *Biochem. J.* **410,** 503–511 (2008).
17. Häcker, U., Nybakken, K. & Perrimon, N. Heparan sulphate proteoglycans: the sweet side of development. *Nature Rev. Mol. Cell Biol.* **6,** 530–541 (2005).
18. Petersen, C. P. & Reddien, P. W. Polarized notum activation at wounds inhibits Wnt function to promote planarian head regeneration. *Science* **332,** 852–855 (2011).
19. Chang, M. V., Chang, J. L., Gangopadhyay, A., Shearer, A. & Cadigan, K. M. Activation of wingless targets requires bipartite recognition of DNA by TCF. *Curr. Biol.* **18,** 1877–1881 (2008).
20. Flowers, G. P., Topczewska, J. M. & Topczewski, J. A zebrafish Notum homolog specifically blocks the Wnt/β-catenin signaling pathway. *Development* **139,** 2416–2425 (2012).
21. Torisu, Y. *et al.* Human homolog of NOTUM, overexpressed in hepatocellular carcinoma, is regulated transcriptionally by β-catenin/TCF. *Cancer Sci.* **99,** 1139–1146 (2008).
22. Baena-López, L. A., Nojima, H. & Vincent, J.-P. Integration of morphogen signalling within the growth regulatory network. *Curr. Opin. Cell Biol.* **24,** 166–172 (2012).
23. Alberts, L. J, *et al. Molecular Biology of the Cell* Ch. 2 (Garland Science, 2008).
24. Basler, K. & Struhl, G. Compartment boundaries and the control of *Drosophila* limb pattern by hedgehog protein. *Nature* **368,** 208–214 (1994).
25. Alexandre, C., Jacinto, A. & Ingham, P. W. Transcriptional activation of hedgehog target genes in *Drosophila* is mediated directly by the cubitus interruptus protein, a member of the GLI family of zinc finger DNA-binding proteins. *Genes Dev.* **10,** 2003–2013 (1996).
26. Teleman, A. A. & Cohen, S. M. Dpp gradient formation in the *Drosophila* wing imaginal disc. *Cell* **103,** 971–980 (2000).
27. Whalen, D. M., Malinauskas, T., Gilbert, R. J. C. & Siebold, C. Structural insights into proteoglycan-shaped Hedgehog signaling. *Proc. Natl Acad. Sci. USA* **110,** 16420–16425 (2013).
28. Belenkaya, T. Y. *et al. Drosophila* Dpp morphogen movement is independent of dynamin-mediated endocytosis but regulated by the glypican members of heparan sulfate proteoglycans. *Cell* **119,** 231–244 (2004).
29. Lum, L. *et al.* Identification of Hedgehog pathway components by RNAi in *Drosophila* cultured cells. *Science* **299,** 2039–2045 (2003).
30. Akiyama, T. *et al.* Dally regulates Dpp morphogen gradient formation by stabilizing Dpp on the cell surface. *Dev. Biol.* **313,** 408–419 (2008).
31. You, J., Belenkaya, T. & Lin, X. Sulfated is a negative feedback regulator of wingless in *Drosophila*. *Dev. Dyn.* **240,** 640–648 (2011).
32. Kirkpatrick, C. A., Dimitroff, B. D., Rawson, J. M. & Selleck, S. B. Spatial regulation of Wingless morphogen distribution and signaling by Dally-like protein. *Dev. Cell* **7,** 513–523 (2004).
33. Baeg, G. H. & Perrimon, N. Functional binding of secreted molecules to heparan sulfate proteoglycans in *Drosophila*. *Curr. Opin. Cell Biol.* **12,** 575–580 (2000).
34. Nardini, M. & Dijkstra, B. W. α/β hydrolase fold enzymes: the family keeps growing. *Curr. Opin. Struct. Biol.* **9,** 732–737 (1999).
35. Krissinel, E. & Henrick, K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr. D* **60,** 2256–2268 (2004).
36. Wu, D. *et al.* Crystal structure of human esterase D: a potential genetic marker of retinoblastoma. *FASEB J.* **23,** 1441–1446 (2009).
37. Duncan, J. A. & Gilman, A. G. A cytoplasmic acyl-protein thioesterase that removes palmitate from G protein α subunits and p21(RAS). *J. Biol. Chem.* **273,** 15830–15837 (1998).
38. Laskowski, R. A., Watson, J. D. & Thornton, J. M. ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res.* **33,** W89–W93 (2005).
39. Orfila, C. *et al.* Expression of mung bean pectin acetyl esterase in potato tubers: effect on acetylation of cell wall polymers and tuber mechanical properties. *Planta* **236,** 185–196 (2012).
40. Xu, Q. *et al.* Vascular development in the retina and inner ear: control by Norrin and Frizzled-4, a high-affinity ligand-receptor pair. *Cell* **116,** 883–895 (2004).
41. Pepinsky, R. B. *et al.* Identification of a palmitic acid-modified form of human Sonic hedgehog. *J. Biol. Chem.* **273,** 14037–14045 (1998).
42. Resh, M. D. Covalent lipid modifications of proteins. *Curr. Biol.* **23,** R431–R435 (2013).
43. Rios-Esteves, J. & Resh, M. D. Stearoyl CoA desaturase is required to produce active, lipid-modified Wnt proteins. *Cell Reports* **4,** 1072–1081 (2013).
44. Reichsman, F., Smith, L. & Cumberledge, S. Glycosaminoglycans can modulate extracellular localization of the wingless protein and promote signal transduction. *J. Cell Biol.* **135,** 819–827 (1996).
45. Fuerer, C., Habib, S. J. & Nusse, R. A Study on the Interactions Between Heparan Sulfate Proteoglycans and Wnt Proteins. *Dev. Dyn.* **239,** 184–190 (2010).
46. Cruciat, C.-M. & Niehrs, C. Secreted and transmembrane wnt inhibitors and activators. *Cold Spring Harb. Perspect. Biol.* **5,** a015081 (2013).
47. de Lau, W., Peng, W. C., Gros, P. & Clevers, H. The R-spondin/Lgr5/Rnf43 module: regulator of Wnt signal strength. *Genes Dev.* **28,** 305–316 (2014).
48. Zebisch, M. *et al.* Structural and molecular basis of ZNRF3/RNF43 transmembrane ubiquitin ligase inhibition by the Wnt agonist R-spondin. *Nature Commun.* **4,** 2787 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank K. Dingwell for supplying purified mWnt3A, H. Bellen for anti-Senseless, C. Alexandre for plasmids and advice, W. Chai for glycosaminoglycan probes, T. Holder for suggestions, T. Malinauskas and C. Lorenz for advice and technical support, T. Walter for technical support with crystallization, W. Lu and Y. Zhao for help with tissue culture, and the organisers of the EMBO Wnt meeting 2012 where our collaboration began. We thank staff at Diamond Light Source beamlines (i02, i03, i04, i04-1, i24) for assistance with data collection (proposal mx8423). This work was supported by the MRC (U117584268 to J.-P.V.; G0900084 to E.Y.J.), the UK Research Council Basic Technology Initiative (Glycoarrays Grant GRS/79268 and EPSRC Translational Grant EP/G037604/1), the Wellcome Trust (Biomedical Resource Grants WT093378MA and WT099197MA) to T.F., the European Union (ERC grant WNTEXPORT; 294523 to J.-P.V., a Marie Curie IEF grant to M.Z.), Cancer Research UK (C375/A10976 to E.Y.J.), and the Japan Society for the Promotion of Science (to S.K.). T.-H.C. was funded by a Nuffield Department of Medicine Prize Studentship in conjunction with Clarendon and Somerville College Scholarships. The Wellcome Trust Centre for Human Genetics is supported by Wellcome Trust Centre grant 090532/Z/09/Z.

Author Contributions Experimental contributions were as follows: *Drosophila* developmental genetics (P.F.L. and S.K.); *Drosophila* cell-based assays (S.K.); human cell-based assays (M.Z. and T.-H.C.); mass spectrometry (S.H., S.K. and A.P.S.); glycan arrays (Y.L., S.K. and T.F.); enzymatic assays (M.Z.); structural biology (M.Z.); peptide synthesis (G.B. and N.O'R.). The project was conceived by S.K., P.F.L., M.Z., E.Y.J. and J.-P.V. The first draft of the paper was written by M.Z., E.Y.J. and J.-P.V. with substantial contributions from P.F.L., S.K. and A.P.S. All authors contributed to the design and interpretation of experiments.

Author Information The crystal structures reported in this paper have been deposited in the Protein Data Bank (PDB) under accession numbers 4WBH, 4UYU, 4UYW, 4UZL, 4UYZ, 4UZ1, 4UZ5, 4UZ6, 4UZ7, 4UZ9, 4UZA, 4UZQ, 4UZJ and 4UZK. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.Z. (matthias.zebisch@evotec.com), E.Y.J. (yvonne@strubi.ox.ac.uk) or J.-P.V. (jvincen@nimr.mrc.ac.uk).

METHODS

Immunostaining and microscopy. The following primary antibodies were used: guinea-pig anti-Senseless (1:1,000, gift from H. Bellen), mouse anti-Patched (1:50, Hybridoma bank), rabbit anti-V5 (1:500, Abcam), mouse anti-V5 (1:500, Invitrogen), rabbit anti-p-Smad3 (1:500, Epitomics), mouse anti-Dlp (1:50, Hybridoma bank), rabbit anti-GFP (1:500, Abcam), mouse anti-Wingless (1:200, Hybridoma bank). Secondary antibodies used were Alexa 488, Alexa 555 and Alexa 647 (1:500, Molecular Probes). Total and extracellular immunostaining of imaginal discs was performed as previously described[49]. Imaginal discs were mounted in Vectashield with 4′,6-diamidino-2-phenylindole (DAPI; Vector Laboratories) and imaged using a Leica SP5 confocal microscope. Confocal images were processed with ImageJ (NIH) and Photoshop CS5.1 (Adobe). All confocal images show a single confocal section. Adult wings were mounted in Euparal (Fisher Scientific) and imaged with a Zeiss Axiophot2 microscope with an Axiocam HRC camera. Adult wing size and L3–L4 intervein distance was measured with ImageJ.

***Drosophila* husbandry and clone induction.** All crosses were performed at 25 °C except those to generate discs shown in Figs 1a, b, 2f and Extended Data Figs 1a, b and 2g, h, k, l in which larvae were reared at 18 °C, the Gal80^{ts} permissive temperature, and then shifted to 29 °C, the restrictive temperature, 16 h before dissection to induce *UAS-notum-V5* expression. To generate mutant clones, larvae were heat-shocked for 1 h at 37 °C at 60 h (±12 h) after egg laying, except for the cross to generate the disc shown in Fig. 2f, which was heat-shocked for 1 h at 37 °C at 84 h (±12 h) after egg laying. Large mutant clones were generated by including a *Minute* mutation on the homologous chromosome[50].

***Drosophila* genotypes.** The following *Drosophila* genotypes were used: *Cyo / UAS-notum-V5 ; tub::Gal80^{ts} / +* (Fig. 1a); *ap-Gal4 / UAS-notum-V5 ; tub::Gal80^{ts} / +* (Fig. 1b); *yw hs-FLP ; notum^{KO} FRT2A / Ubi::GFP M FRT2A* (Fig. 1c); *yw hs-FLP ; dally^{MH32} dlp^{MH20} FRT2A / Ubi::GFP M FRT2A* (Fig. 1d); *UAS-notum-V5 / + ; dpp-Gal4 / +* (Fig. 2a); *UAS-notum-V5 / + ; dpp-Gal4 dlp^{MH20} / dlp^{MH20} FRT2A* (Fig. 2b); *UAS-notum-V5 / + ; dpp-Gal4 dally^{MH32} / dally^{MH32} FRT2A* (Fig. 2c); *UAS-notum-V5 / tub::dlp-CD8 ; dpp-Gal4 dlp^{MH20} / dlp^{MH20} FRT2A* (Fig. 2d); *lex-OP-notum-V5 / en-Gal4 UAS::GFP ; dpp-lexA / UAS-sulf-RNAi* (Fig. 2e); *yw hs-FLP / tub::Gal80^{ts} ; UAS-notum-V5 / ptc-Gal4 UAS::GFP ; dally^{MH32} dlp^{MH20} FRT2A / Ubi::GFP FRT2A* (Fig. 2f); *Cyo / UAS-notum-V5 ; tub::Gal80^{ts} / +* (Extended Data Fig. 1a); *ap-Gal4 / UAS-notum-V5 ; tub::Gal80^{ts} / +* (Extended Data Fig. 1b); *yw hs-FLP ; notum^{KO} FRT2A / Ubi::GFP M FRT2A* (Extended Data Fig. 1c); *notum^{141} Ubi::GFP FRT2A / Mkrs* (Extended Data Fig. 1d); *notum^{KO} FRT2A / notum^{141} Ubi::GFP FRT2A* (Extended Data Fig. 1e); *UAS-NRT-Wg / + ; dpp-Gal4 / +* (Extended Data Fig. 2a); *UAS-NRT-Wg / + ; dpp-Gal4 / UAS-notum* (Extended Data Fig. 2b); *dally-GFP* (protein-trap, DGRC 115-064) (Extended Data Fig. 2e, f, i, j); *Cyo / UAS-notum-V5 ; tub::Gal80^{ts} / +* (Extended Data Fig. 2g); *ap-Gal4 / UAS-notum-V5 ; tub::Gal80^{ts} / +* (Extended Data Fig. 2h); *Cyo / UAS-notum-V5; tub::Gal80^{ts} / dally-GFP* (Extended Data Fig. 2k); *ap-Gal4 / UAS-notum-V5; tub::Gal80^{ts} / dally-GFP* (Extended Data Fig. 2l); *UAS-NRT-wg / + ; UAS-notum dally^{MH32} / dpp-Gal4 dally^{MH32}* (Extended Data Fig. 3a); *UAS-notum-V5 / UAS-notum-V5* (Extended Data Fig. 3b); *sal-Gal4 / UAS-notum-V5* (Extended Data Fig. 3c); *sal-Gal4 / UAS-notum-V5 ; dally^{MH32} FRT2A / dally^{MH32} FRT2A* (Extended Data Fig. 3d).

Generation of *notum* knockout by homologous recombination. *notum^{KO}* was generated by homologous recombination using reagents and crossing schemes described previously[51]. The homology arms were amplified from *w^{1118}* genomic DNA. The primers 5′-GATCGCTAGCCGAGAAAGACACAAACGAAGATC AAC-3′ and 5′-GATCGGTACCCGATTCGATTACACATAGATATAGAATA G-3′ were used to amplify the upstream 5-kilobase (kb) homology arm, which was cloned into pTV as an Nhe1-Kpn1 fragment. The primers 5′-GATCACT AGTGTTATCAAAAGCGAACGCCGCAATAC-3′ and 5′-GATCAGATCTCT GGAATTGATTTGATTCGATTGCGGTG-3′ were used to amplify the downstream 3-kb homology arm, which was cloned into pTV as a SpeI-BglII fragment. *notum^{KO}* deletes 82-base pair (bp) sequence of the first exon that encodes the signal sequence. As expected, *notum^{KO}* behaved as a null. It was recombined onto *FRT2A* for clonal analysis.

Transgene to express Dlp-CD8. Dlp-CD8 was made by replacing the C terminus where the GPI anchor is normally added with mouse CD8 transmembrane domain and GFP. The primers 5′-GATGAATTCGGCGCGCCATGCTACATCAGCAG CAACAAC-3′ and 5′-GCATGCGGCCGCCTCGATTGTCATTGGCCCCG-3′ were used to amplify 2,193 bp of the cDNA encoding a polypeptide lacking Asp 734, where GPI is normally appended. This fragment was cloned in frame as an EcoR1-Not1 fragment in UAS-HRP-CD8-GFP (deleting horseradish peroxidase (HRP)) and the Dlp-CD8-encoding fragment was then transferred to pMTV5 as an EcoR1-Xho1 fragment. From there it was transferred to pTubulin as a Kpn1-Mlu1 fragment. This transgene rescued viability and wing patterning in *dlp* mutant homozygotes, which otherwise do not survive beyond pupal stages, a strong indication that Dlp does not need to be GPI anchored for normal development.

Expression vectors for cultured *Drosophila* cells. *Drosophila* S2 or S2R+ (*Drosophila* Genomics Resource Centre, DGRC), were cultured at 25 °C in Schneider's medium plus L-glutamine (Sigma) containing 10% (v/v) fetal FBS (Life Technologies) and 0.1 mg ml^{-1} pen/strep (Life Technologies). To generate plasmids expressing V5-tagged dNotum, the dNotum cDNA (from S. Cohen) was amplified, adding a V5 tag (GKPIPNPLLGLDST) at the C terminus. This fragment was then inserted into pActin, pUAST or pLotattB[52] to generate pAct-Notum-V5, pUAST-Notum-V5 or pLotattB-Notum-V5, respectively. A stable S2 line expressing V5-tagged dNotum (S2 act-Notum-V5) was generated by transfection of S2 cells with pAct-Notum-V5 and pCoHygro (Invitrogen) followed by drug selection. Wingless was expressed from pTub-Wg, which was prepared by inserting the Wg cDNA from pKS-Wg[53] into pTubulin. HA-tagged Dally was expressed from pAct-Dally-HA, prepared by inserting Dally-HA excised from pMT-Dally-HA (from S. Cohen) into pActin. To conveniently manipulate the coding sequence of Dlp, three nucleotides (GTC) were inserted at positions 2100–2102 (nucleotide numbering with the A of first codon at position 1) to introduce a SalI site in KS-Dlp. This was used to insert DNA encoding an HA tag flanked by Glycine (GYPYDVPDYAG) and thus generate pKS-Dlp-HA. The Dlp-HA was then inserted into pTubulin to make pTub-Dlp-HA.

PIPLC treatment of imaginal discs and cultured cells. Wing imaginal discs were treated with PIPLC as previously described in[54] with some modifications. In brief, discs were dissected from third instar larvae and incubated in Schneider's medium with 10% FBS containing 10 U ml^{-1} of PIPLC (Molecular Probes) at room temperature for 30 min. After treatment, the discs were washed three times with Schneider's medium before extracellular staining (no detergent). S2 cells transfected with pTub-Dlp-HA and pActin (mock), or pTub-Dlp-HA and pAct-Notum-V5 as well as the corresponding conditioned medium were collected (total volume 300 µl) and treated with PIPLC (final concentration 1 U ml^{-1}) for 1.5 h at 25 °C before phase separation.

Phase separation assay. The phase separation assay was performed as previously described with some modifications[55]. After PIPLC treatment, 200 µl of pre-condensed Triton X-114 (Sigma) was added to the reaction mixtures (Triton X-114 final concentration ~2%). The extracts were incubated for 15 min on ice and then centrifuged at 10,000*g* for 10 min at 4 °C. The supernatant were transferred to new tubes and warmed at 37 °C in a water bath for 10 min. After a second centrifugation (10,000*g* for 10 min at room temperature), the upper phases (aqueous) and lower phases (detergent) were collected separately and mixed with 4× sample buffer (Life Technologies) for analysis by immunoblotting.

Immunoblotting. Samples were run on 4–12% Bis-Tris NuPAGE gels (Invitrogen) with MOPS buffer. Proteins on gel were transferred onto nitrocellulose membrane using iBlot gel transfer System (Invitrogen). The membranes were washed with dH_2O and blocked with 5% skimmed milk in 0.1% Tween-20 PBS (PBS-T) for 30 min at room temperature. Membranes were incubated with primary antibodies (mouse monoclonal anti-V5; Life Technologies, 1:5,000 and rat anti-HA; Roche, 1:2,500) diluted in 5% milk PBS-T overnight at 4 °C and washed with PBS-T three times before incubation with HRP-conjugated secondary antibodies (anti-mouse or anti-rat; Biorad, 1:5,000). Membranes were washed again in PBS-T, developed using ECL prime western blotting detection system (GE Healthcare) and exposed to film.

Glycan array. CM obtained from S2 cells expressing V5-tagged dNotum was overlaid on a focused neoglycolipid-based glycan array containing lipid-linked GAG oligosaccharide probes (see http://www1.imperial.ac.uk/glycosciences/ and refs 56, 57) and allowed to bind for 90 min. The array was then washed and stained with anti-V5 mouse monoclonal antibody (Invitrogen) followed by biotinylated anti-mouse IgG (Sigma). Binding was detected with Alexa Fluor 647-labelled streptavidin. Fluorescence intensity was quantified and data analysis was performed with dedicated microarray software. No binding was observed when control medium was used instead of the conditioned medium or when the anti-V5 was used in the absence of the dNotum medium (data not shown).

Large-scale expression of Notum constructs. The cDNA coding for mature hNOTUM (residues Arg 38–Ser 496) was cloned into the pHLsec vector[58] that adds a C-terminal His6- or His10-tag. After the crystal structure was solved in crystal forms I and II (see below) and the folded region identified, a shorter construct hNOTUM$_{core}$ comprising Ser 81–Thr 451, Cys330Ser, was found to provide higher expression levels, thanks in part to the removal of the non-conserved Cys 330, which provides a free, surface-exposed sulfhydryl. Expression of wild-type protein resulted in non-quantitative spontaneous crosslinking of the protein, a problem that was not observed with the Cys330Ser variant.

For dNotum, we initially attempted to express Asp83–Thr617. However, a large unstructured and non-conserved domain of 22 kilodaltons (kDa) (Arg 416—Lys 597) was found to interfere with crystallization. This domain, which was not present in hNOTUM, was deleted and replaced by GNNNG to generate dNotum$_{\Delta loop}$. Note that this domain could provide an additional glypican-binding surface since it is

highly basic (pI = 12.4). Proteins were transiently expressed in HEK293T cells and purified as described[48]. Proteins for crystallization were expressed either in GntI-deficient HEK293S cells or in HEK293 cells treated with kifunensine (1 mg l^{-1}). Before crystallization the proteins were treated with endoglycosidase F1 at a ratio of 1:100. Proteins intended for kinetic studies were stored in 10 mM Tris-HCl, pH 7.5, 1 mM EDTA, 50 mM NaCl and 50% (v/v) glycerol at −20 °C.

SPR equilibrium binding studies. Affinity between variants of hNOTUM and GPC3 or sulfated GAG was measured at 25 °C in 10 mM HEPES/NaOH, pH 7.5, 150 mM NaCl and 0.005% Tween20 using a Biacore T200 machine (GE Healthcare). GPC3 constructs (see below) or sulfated GAGs were coupled to a streptavidin-coated sensor chip via a biotin label and purified Notum proteins were used as analyte. Biotinylated GAGs were produced as described[59]. To produce biotinylated GPC3 we proceeded as follows. The cDNA encoding human GPC3 (full-length except for the lack of endogenous signal sequence, Pro 31—Asn 538) or GPC3$_{\Delta GAG}$ (lacking a C-terminal stretch that normally contains the GAG attachment sites, Pro 31–Phe 493) was cloned into a variant of the pHLsec vector, which introduces a recognition sequence for the *Escherichia coli* BirA enzyme at the C terminus. Biotinylation at this sequence tag was performed by co-transfection of HEK293T cells with the GPC3 construct and an *E. coli* BirA expression construct. The synthetic BirA gene was codon-optimized and carried a C-terminal KDEL-tag for retention in the endoplasmic reticulum. The BirA plasmid was used at 20% of total DNA. The expression medium was supplemented with 100 µM of sterile biotin prepared as a 2 mM stock in PBS. After 3 days, the conditioned medium was cleared from cell debris and repeatedly buffer-exchanged to remove free biotin. The chip surface was precoupled with approximately 10,000 resonance units (RU) of streptavidin. Approximately 500 RU of GPC3 was immobilized. The amount of immobilized GAGs could not be measured. After each injection of analyte the chip surface was regenerated with 1 M NaCl, 10 mM HEPES/NaOH, pH 7.5, to return to baseline levels. Data were fit to a Langmuir adsorption model ($B = B_{max}C/(K_d + C)$, where B is the amount of bound analyte and C is the concentration of analyte in the sample. Data were then normalized to a maximum analyte binding value of 100. For the design of heparin binding site mutants, the following considerations were taken into account. If, based on the crystal structure, the hydrophobic part of the side chain (for example, Arg, Lys, His) was estimated to be of no structural importance, then the residue was mutated to serine. In all other cases it was mutated to glutamine (Arg, Lys) or asparagine (His) to keep the overall structure as native as possible.

Heparin affinity chromatography. We compared the affinity of hNOTUM variants for heparin using a 1 ml HiTrap Heparin HP column (GE Healthcare). The column was equilibrated in 10 mM Tris-HCl, pH 8.0. Sample protein (120 µg), diluted into binding buffer, was injected onto the column. After washing of the column with five column volumes of binding buffer the protein was eluted in a linear gradient to 1.0 M NaCl over 20 column volumes. The flow rate was 2 ml min^{-1}.

Chromogenic Notum activity assays. Steady-state carboxylesterase activity measurements of hNOTUM were performed in 50 mM MES/NaOH, 100 mM NaCl, pH 6.5, using different chromogenic pNP esters (Sigma; number indicates carboxylic acid chain length). Substrate stocks in DMSO were adjusted to concentrations between 20 mM (pNP16) and 2 M (pNP2) and diluted into reaction buffer. In tests using the short and soluble substrates pNP2 and pNP4, the final DMSO concentration was only 0.1%. In tests using longer pNP substrates or in comparative studies the final DMSO and Triton concentration was kept constant at 2.5% (v/v) and 0.5% (w/v) respectively. The required amount of a 20-mM substrate stock was first mixed 1:1 with a 20% (w/v) solution of Triton X-100 in reaction buffer. The resulting emulsion was then diluted with reaction buffer and vigorously agitated to avoid precipitation. Reactions were started by addition of 5–10 µl of protein at concentrations between 0.1 and 4 mg ml^{-1}. Substrate was measured using a Varian Cary 50 spectrophotometer by following the absorption change at 405 nm. The extinction coefficient of *p*-nitrophenol in reaction buffer was established to be 4,070 M^{-1} cm^{-1}. Although Triton X-100 was required to maintain the solubility of long ester substrates and fatty acids, it was itself an inhibitor of Notum (Extended Data Fig. 8a, b). We assume that the hydrophobic region of Triton X-100 has a propensity to bind to the active site pocket. This notion is supported by the observation that the much larger sterol-based detergent CHAPS evokes no inhibition. On the basis of this assumption of competitive inhibition by Triton X-100 we determined its inhibition constant to be 466 µM and used it to calculate a corrected Michaelis constant for pNP8 turnover.

Cell-based Notum activity assays. To assay Wingless signalling in *Drosophila* cells, a modified TOPFlash vector called WISIR, comprising a TCF-responsive promoter driving Firefly luciferase and a ubiquitous promoter driving Renilla luciferase[60] was used. To assess the repressive activity of dNotum, S2R+ cells were transfected separately in 6-well plates with pTub-Wg (2 µg), pAct-Notum-V5 (2 µg) and WISIR (0.3 µg). The transfected cells were then cultured for 2 days at 25 °C and then mixed in equal ratio. Firefly and Renilla luciferase levels were measured 24 h later with Dual-Glo luciferase reporter assay system. As controls, cells transfected with WISIR alone or with WISIR and pTub-Wg were mixed with mocked treated cells. Firefly luciferase activity was normalized to *Renilla* luciferase activity and the average of triplicate samples was calculated.

hNOTUM inhibition of Wnt signalling in mammalian cells was assessed in stably transfected SuperTopFlash (STF) HEK293 cells[40]. These were treated with conditioned medium from Wnt3A-producing L cells[4] with or without recombinant purified hNOTUM. To reveal the direct action of hNOTUM on Wnt we proceeded as follows. Wnt3A CM was dialysed for 24 h against ten volumes of tissue-culture-grade PBS and then sterile-filtered with the aim to remove chelators that might interfere with TALON-binding (see below). To 500 µl of such dialysed Wnt3A, 5 µl hNOTUM protein or an unrelated control protein (mock) at a concentration of 1 mg ml^{-1} was added and incubated for the indicated time at room temperature (23 °C). To stop the enzymatic reaction we added 50 µl of fresh 50% slurry of cobalt affinity beads (TALON resin) equilibrated against tissue culture grade PBS and 5 µl of 500 mM imidazole in PBS. After vigorous shaking the solution was incubated for 1 h at room temperature on a vertical rotator. Beads containing the His_{10}-tagged hNOTUM protein were removed by centrifugation (3,000*g*, 5 min) and discarded. The supernatant was cleared again by centrifugation at maximum speed (16,000*g*, 5 min). 100 µl of the reaction solution was then added to STF cells seeded the previous day at 50,000 cells per 100 µl and per well into 96-well plates. The Wnt-induced luciferase activity was measured after 16–20 h using the Glo kit (Promega) and an Ascent Luminoskan luminometer (Labsystems) following the instructions of the manufacturer. Data represent average of quadruplicate measurements ± s.d. The incubation time with cells was kept constant for all compared samples.

To assess hNOTUM inhibition of Norrin signalling, STF cells seeded in 96-well plates were transfected after 24 h with 200 ng DNA: 60 ng each of hFZ4 and hLRP6 plasmids, 30 ng of Tspan-12 plasmid, and 50 ng constitutive *Renilla* luciferase plasmid (pRL-TK from Promega). Cells were stimulated 24 h after transfection with 10 µg ml^{-1} recombinant Norrin (T.-H. Chang *et al.* manuscript in preparation), which had been preincubated for 24 h with 10 µg ml^{-1} hNOTUM variants or FCS as a control. Firefly and *Renilla* luciferase activities were measured 48 h later with Dual-Glo luciferase reporter assay system (Promega). Firefly luciferase activity was normalized to *Renilla* luciferase activity and the average of triplicate samples was calculated.

Crystallization, data collection and structure determination. Concentrated proteins were subjected to sitting drop vapour diffusion crystallization trials employing a Cartesian Technologies pipetting robot and typically consisted of 100–300 nl of protein solution and 100 nl of reservoir solution. A detailed discussion of the multiple conditions in which crystal growth occurred is provided in Supplementary Information. Standard methods were used for X-ray diffraction data collection and structure determination, distinctive details for the series of crystal structures are discussed in Supplementary Information.

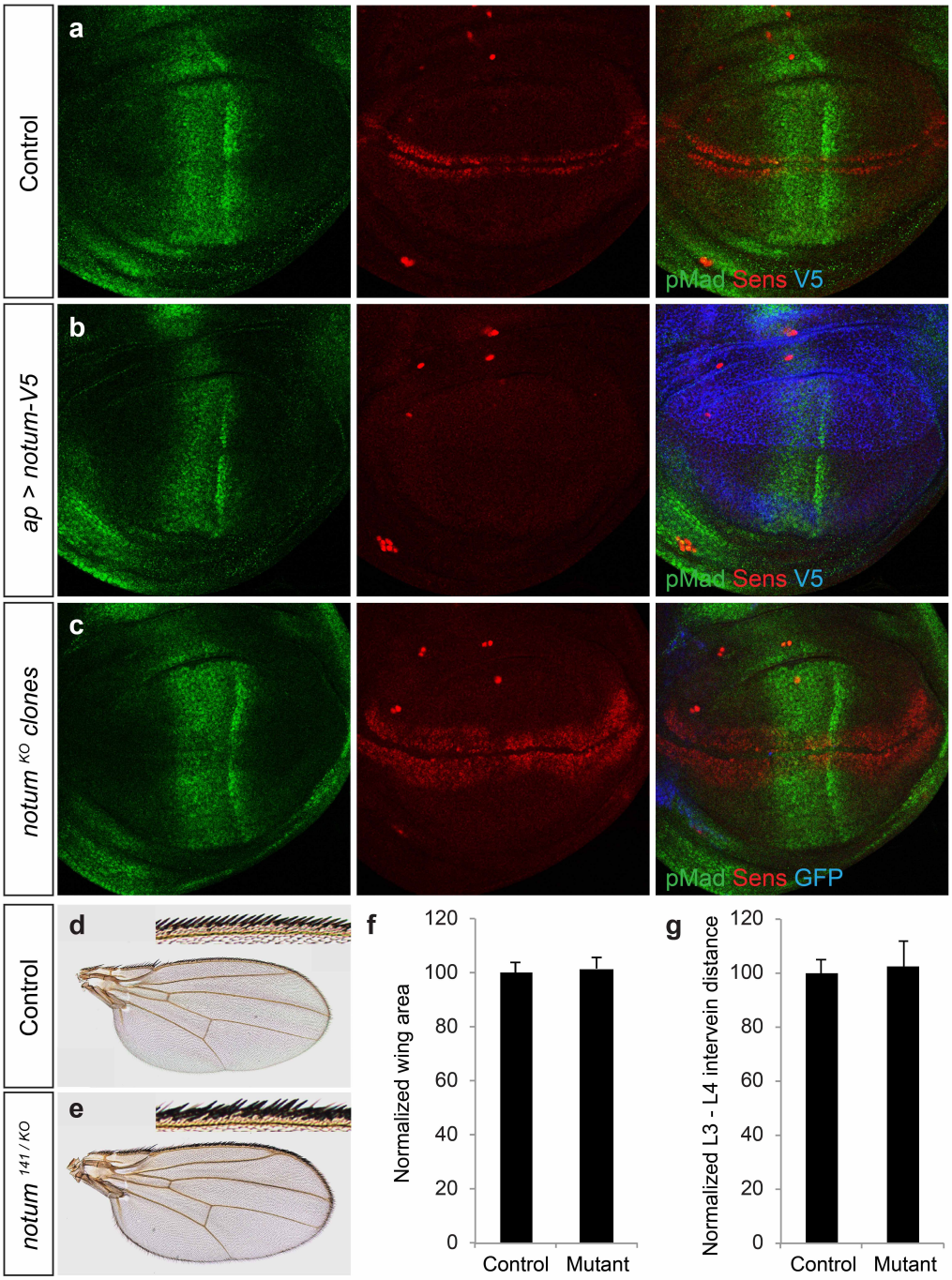
Mass spectrometric analysis of the effect of Notum on Wnt3A protein. As a general method to quantify the levels of delipidated Wnt3A protein by LC–MS-MS, we used an isotope coded alkylation reaction targeting cysteines to multiplex mass spectrometry signal. One sample was reacted with heavy iodoacetamide (IAA, $^{13}C_2D_2HINO$) and the other with the light version (C_2H_3INO) and consequently, peptide signal doublets appeared at Δ4D per cysteine with peak areas used for relative quantification. Wnt3A protein (500 ng, purified from L cell conditioned medium by K. Dingwell (NIMR) according to ref. 4) was mixed with purified hNOTUM (1 µl of purified $hNOTUM_{core}$ at 25 ng $µl^{-1}$) in the following buffer: 20 mM Tris-HCl buffer (pH 7.5) containing 500 mM NaCl, 0.5 mM EDTA, 0.5% CHAPS and 5% glycerol and left together for 16 h at 25 °C. The reaction was quenched by addition of 4× LDS sample buffer (Life Technologies). Coomassie blue stained bands from SDS–PAGE were excised from the gel and cut in half and destained by incubating for 45 mins with 200 mM ammonium bicarbonate (ABC)/60% acetonitrile (ACN). To reduce cysteines, buffer was refreshed with the inclusion of 10 mM dithiothreitol (DTT) for 15 min. After washing, half of the gel pieces were incubated in 20 mM heavy or light IAA in 100 mM ABC/60% ACN buffer in the dark for 30 min. Proteins were digested using a 4 h in-gel trypsin digestion step in 100 mM ABC and then quenched with 0.1% TFA. Equal aliquots of heavy and light reaction were mixed to generate forward and reverse labelled samples. Duplicate LC–MS analysis was performed using an Ultimate3000 RSLC system coupled to a LTQ-Orbitrap Velos-Pro mass spectrometer (Thermo Scientific). The instrument was operated in an alternating targeted MS/MS and data dependent acquisition mode. CHGLSGSCEVK and the control peptide AGIQECQHQFR were targeted for MS/MS. MS/MS spectra were searched using Mascot v2.3 and identifications imported as a spectral library into Skyline software v2.6.0.6709. Skyline was used for peaks extraction and areas determination

Mass spectrometric analysis of Wnt3A and Shh peptides. Delipidation assays were performed by reacting 3 µg of synthetic peptides (synthesis described in Supplementary Information) with 1 µl of enzyme ($hNOTUM_{core}$ or $hNOTUM_{core}$

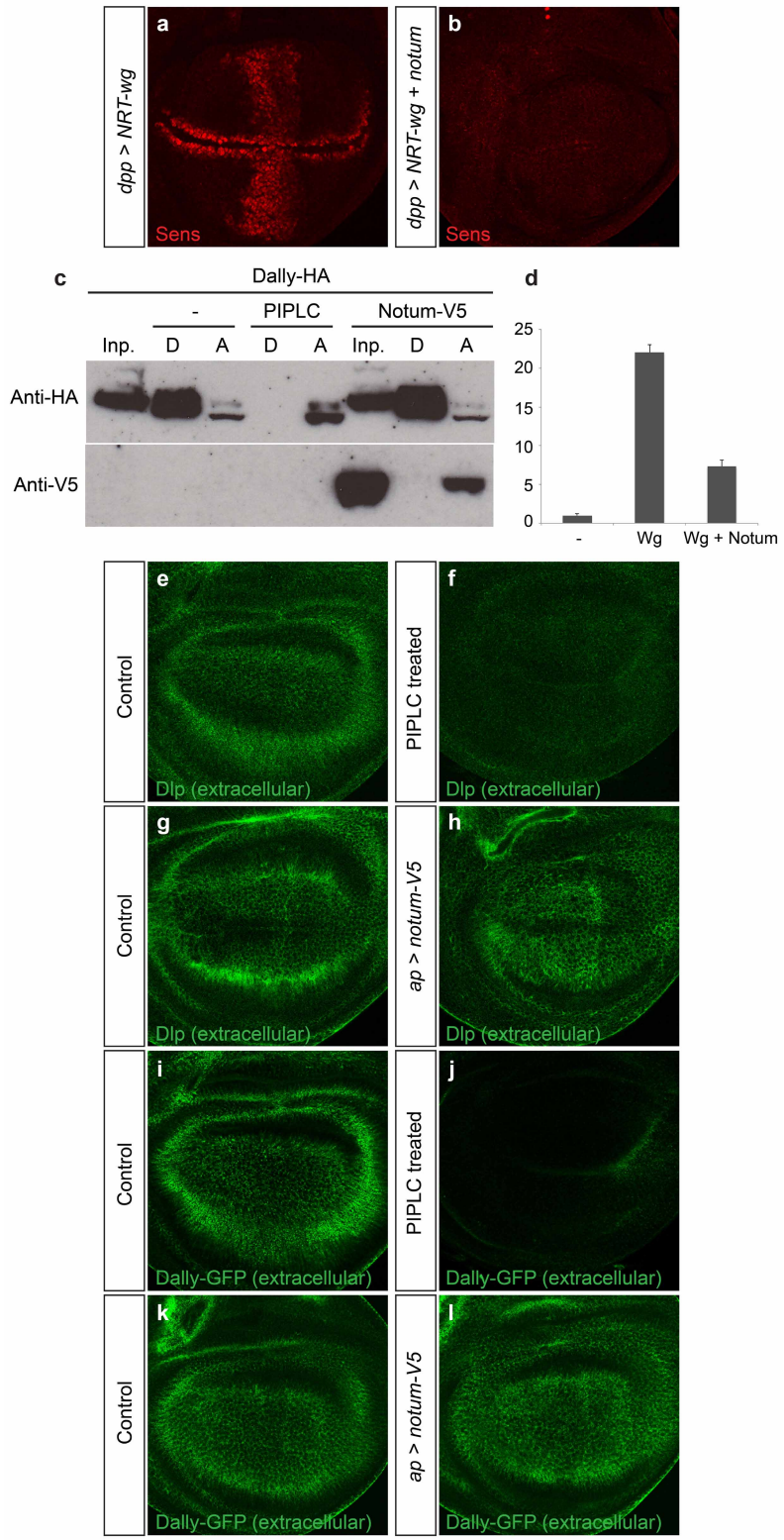
(S232A); 25 ng μl^{-1}) in 20 mM ammonium bicarbonate buffer (total volume 5 µl) for 16 h at 25 °C. The reaction was quenched with 0.1% TFA and samples were desalted using c18 zip tips. Samples were prepared in α-cyano-4-hydroxycinnamic acid in 50:50 water/acetonitrile with 0.1% TFA. MALDI-TOF spectra were acquired using an ABSCIEX 5800 TOF/TOF systems and analysed using data explorer v4.11.

Statistics. No statistical methods were used to predetermine sample size.

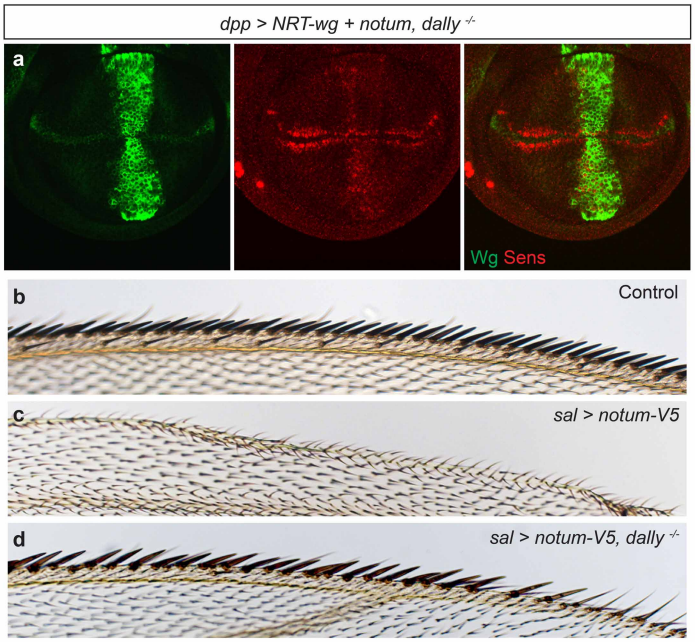
49. Beckett, K. *et al. Drosophila* S2 cells secrete wingless on exosome-like vesicles but the wingless gradient forms independently of exosomes. *Traffic* **14,** 82–96 (2013).
50. Vincent, J.-P., Kolahgar, G., Gagliardi, M. & Piddini, E. Steep differences in wingless signaling trigger myc-independent competitive cell interactions. *Dev. Cell* **21,** 366–374 (2011).
51. Baena-Lopez, L. A., Alexandre, C., Mitchell, A., Pasakarnis, L. & Vincent, J. P. Accelerated homologous recombination and subsequent genome modification in *Drosophila*. *Development* **140,** 4818–4825 (2013).
52. Yagi, R., Mayer, F. & Basler, K. Refined LexA transactivators and their use in combination with the *Drosophila* Gal4 system. *Proc. Natl Acad. Sci. USA* **107,** 16166–16171 (2010).
53. Alexandre, C., Baena-Lopez, A. & Vincent, J.-P. Patterning and growth control by membrane-tethered Wingless. *Nature* **505,** 180–185 (2014).
54. Marois, E., Mahmoud, A. & Eaton, S. The endocytic pathway and formation of the Wingless morphogen gradient. *Development* **133,** 307–317 (2006).
55. Doering, T. L., Englund, P. T. & Hart, G. W. Detection of glycophospholipid anchors on proteins. *Curr. Prot. Prot. Sci.* **Chapter 12,** Unit 12.15 (2001).
56. Fukui, S., Feizi, T., Galustian, C., Lawson, A. M. & Chai, W. Oligosaccharide microarrays for high-throughput detection and specificity assignments of carbohydrate-protein interactions. *Nature Biotechnol.* **20,** 1011–1017 (2002).
57. Palma, A. S., Feizi, T., Childs, R. A., Chai, W. & Liu, Y. The neoglycolipid (NGL)-based oligosaccharide microarray system poised to decipher the meta-glycome. *Curr. Opin. Chem. Biol.* **18,** 87–94 (2014).
58. Aricescu, A. R., Lu, W. & Jones, E. Y. A time- and cost-efficient system for high-level protein production in mammalian cells. *Acta Crystallogr. D* **62,** 1243–1250 (2006).
59. Malinauskas, T., Aricescu, A. R., Lu, W., Siebold, C. & Jones, E. Y. Modular mechanism of Wnt signaling inhibition by Wnt inhibitory factor 1. *Nature Struct. Mol. Biol.* **18,** 886–893 (2011).
60. Gagliardi, M., Hernandez, A., McGough, I. J. & Vincent, J.-P. Inhibitors of endocytosis prevent Wnt/Wingless signalling by reducing the level of basal β-Catenin/Armadillo. *J. Cell Sci.* **127,** 4918–4926 (2014).
61. Glise, B. *et al.* Shifted, the *Drosophila* ortholog of Wnt inhibitory factor-1, controls the distribution and movement of Hedgehog. *Dev. Cell* **8,** 255–266 (2005).



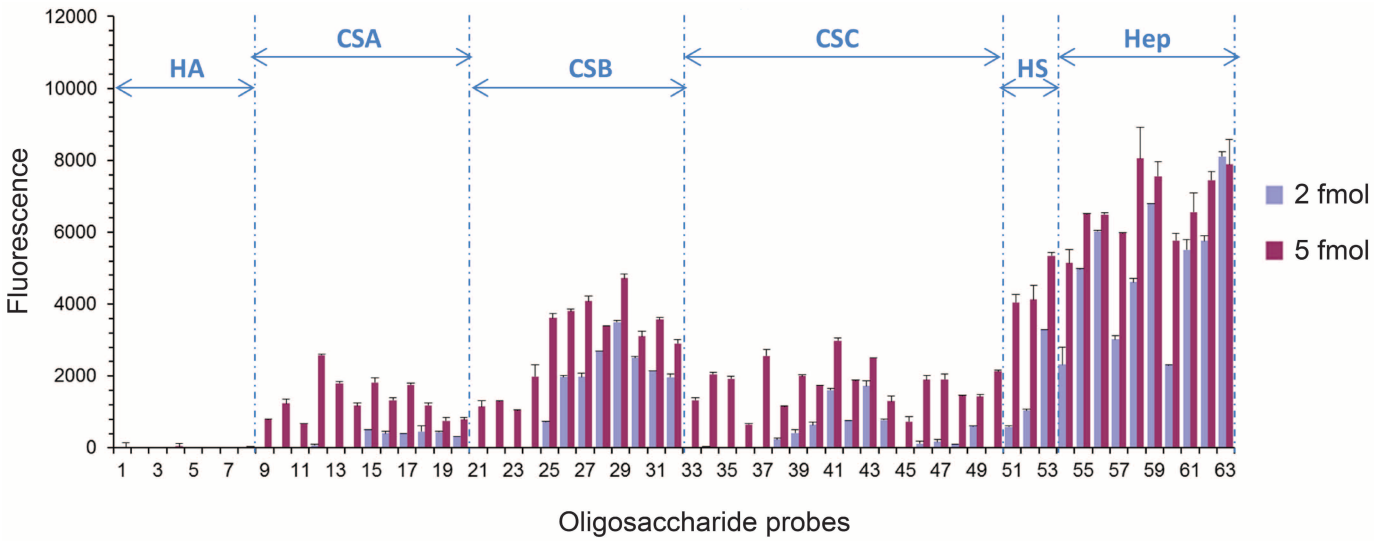
Extended Data Figure 1 | Notum modulates Wingless, but not Dpp or Hedgehog signalling. **a**, **b**, Overexpression of dNotum-V5 with the *apterous-Gal4* driver, which is expressed in the dorsal compartment, prevents expression of Senseless (Sens) (**b**, middle), a Wingless target gene, but has little effect on phospho-Mad (pMad) immunoreactivity (**b**), an indicator of Dpp signalling. **c**, Loss of *notum* activity, achieved by generating large patches of $notum^{KO}$ tissue (see Methods), marked by the loss of GFP, leads to broadening of Senseless expression but does not affect pMad immunoreactivity. **d**–**g**, Strong, but not complete, reduction of *notum* activity led to ectopic wing margin bristles (compare insets in **d** and **e**) but had no significant effect on wing area, which is sensitive to Dpp signalling (**f**) ($P = 0.26$, Student's *t*-test), or on the distance between L3 and L4 veins, which is affected by changes in Hedgehog signalling[61] (**g**) ($P = 0.41$, Student's *t*-test). In total, 19 control ($notum^{141/+}$) and 17 mutant ($notum^{141/KO}$) wings were analysed. Error bars in **f** and **g** are s.d.



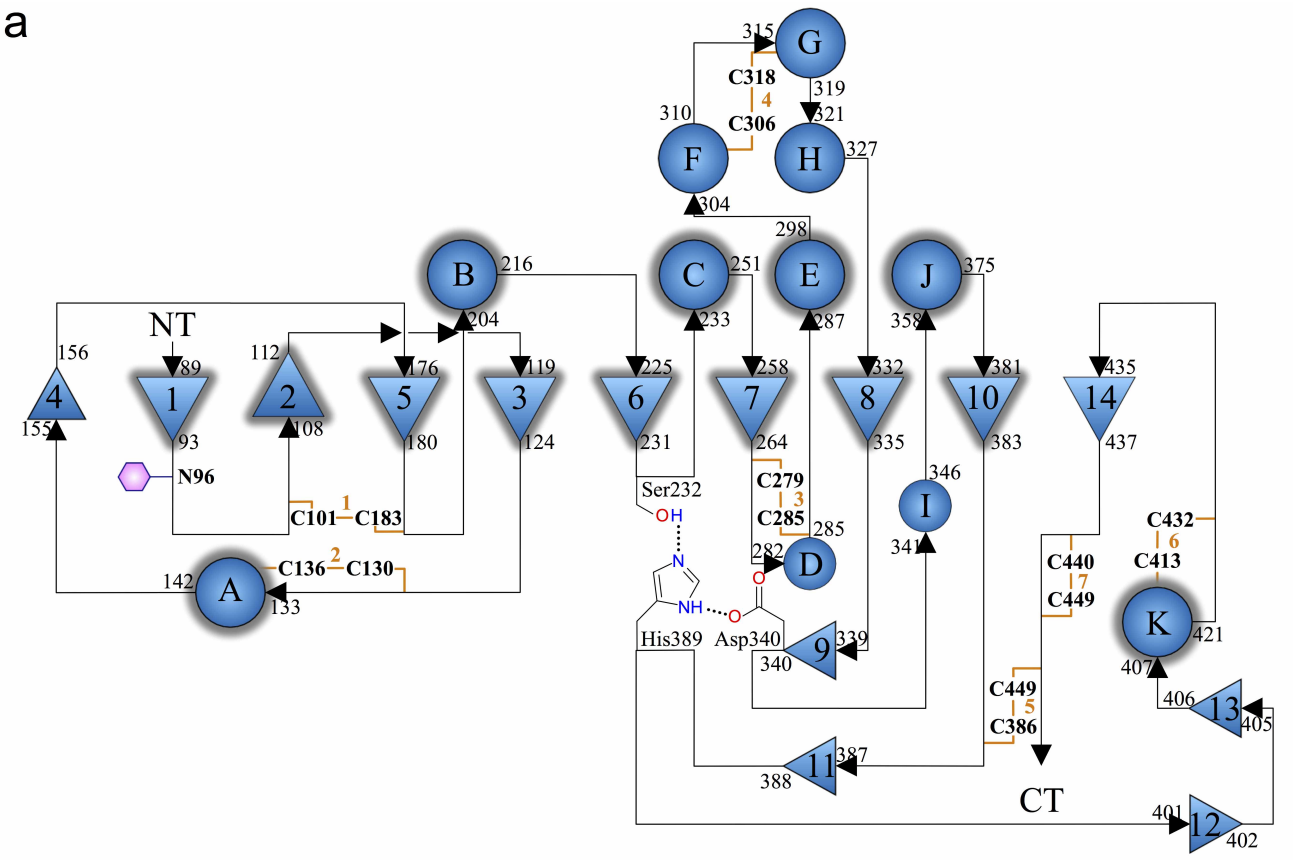
Extended Data Figure 2 | dNotum does not cleave the GPI anchor of glypicans. **a**, **b**, Ectopic expression of Senseless caused by *NRT-wingless*, as well as endogenous Senseless, is suppressed by co-expression of dNotum. *NRT-wingless* and *notum* are expressed in a vertical band under the control of *dpp-Gal4*. **c**, Western blot analysis of phase-separated extracts of S2 cells transfected with a plasmid expressing HA-tagged Dally. In control extracts, Dally is found largely in the detergent (D) phase. Coexpression of dNotum–V5 from a plasmid had no effect, while treatment with PIPLC shifted all detectable Dally to the aqueous (A) phase. **d**, dNotum-V5 expression as in **c** was sufficient to suppress Wingless-induced TOPFlash activity. Cells were transfected with a dual luciferase TOPFlash reporter[60] along with a mock plasmid (−), *tubulin::wingless* (Wg), or *tubulin::wingless* + *actin::notum-V5* (Wg + Notum). **e**–**h**, Extracellular Dlp in control (**e**, **g**), PIPLC-treated (**f**) or *apterous-Gal4 UAS-notum-V5* (**h**) imaginal discs. **i**–**l**, Extracellular anti-GFP staining of imaginal discs from gene trap line expressing Dally–GFP fusion protein. Discs were treated with a mock solution (**i**) or PIPLC (**j**) (same discs as in **e** or **f**, respectively, but here showing Dally protein). In a separate experiment, dNotum was overexpressed with *apterous-Gal4* in the *Dally-GFP* background (**l**). No change in the distribution of extracellular GFP could be seen compared to that in control discs (**k**, no *apterous-Gal4*).



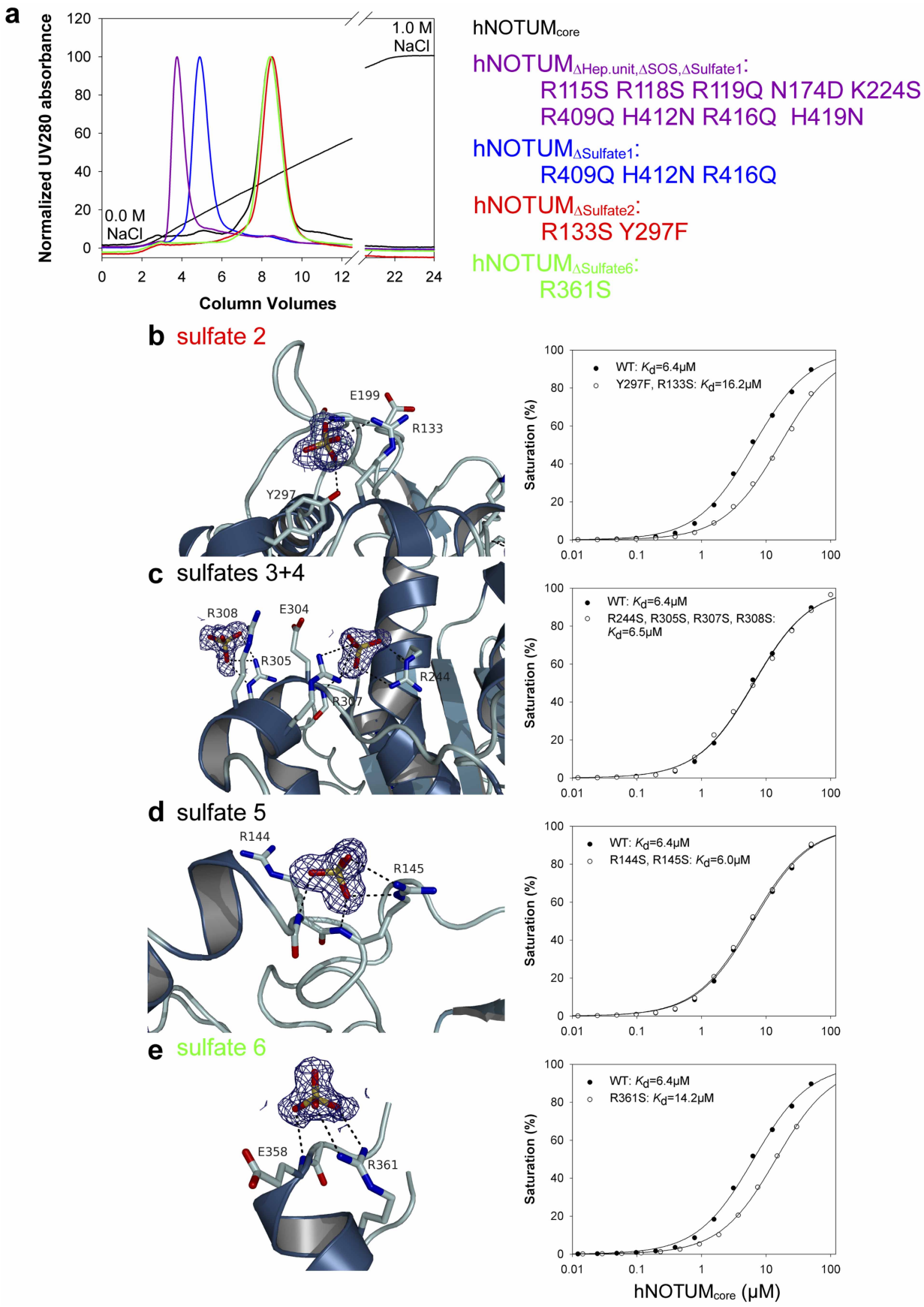
Extended Data Figure 3 | dNotum requires Dally to inhibit Wingless signalling. **a**, Wingless and Senseless expression in a *dally*$^{-/-}$ wing imaginal disc expressing *NRT-wingless* and *notum* under the control of *dpp-Gal4*. Some *senseless* expression remains, indicating that, in the absence of Dally, dNotum is a poor inhibitor of NRT-Wingless-induced (as well as endogenous) signalling. **b–d**, Anterior margin of wings from control, *spalt (sal)-Gal4 UAS-notum-V5*, and *sal-Gal4 UAS-notum-V5 dally*$^{-/-}$ animals. Removal of *dally* rescues the loss of margin bristles caused by dNotum overexpression.



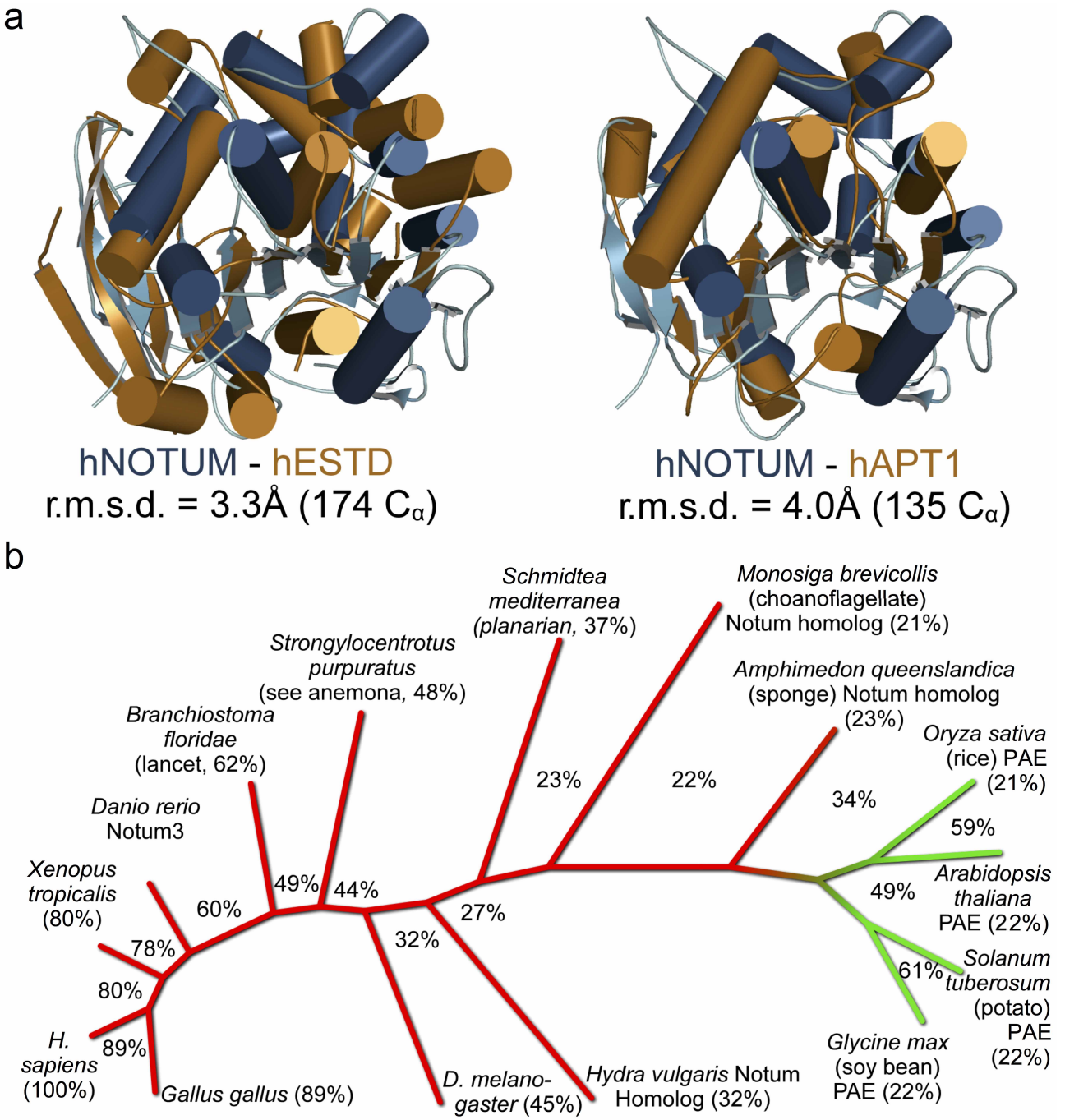
Extended Data Figure 4 | dNotum binds to sulfated glycans. Binding of dNotum-V5 to a GAG oligosaccharide array, detected by immunofluorescence. CSA/B/C, chondroitin sulfate A/B/C; HA, hyaluronic acid, hep, heparin; HS, heparan sulfate. Details on the array are provided in the Methods.



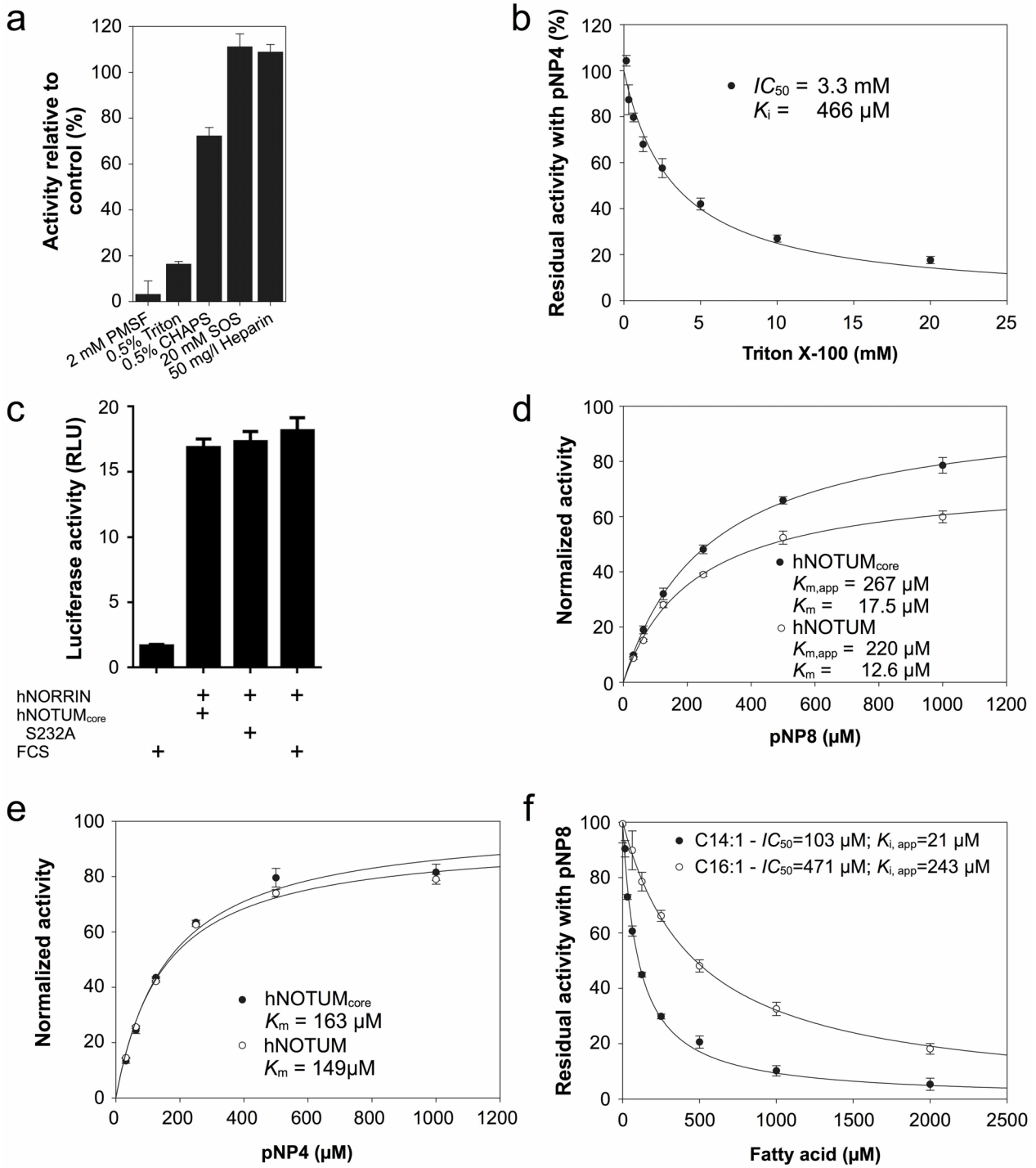
Extended Data Figure 5 | Additional structural information on Notum. **a**, Topology plot of hNOTUM. β-strands are shown as numbered triangles and α-helices as circles labelled in alphabetical order from the N to C terminus (NT to CT). Structural elements conserved among most α/β-hydrolases are outlined in grey. **b**, Comparison of the two most conformationally distinct hNOTUM structures (from crystal forms III and V). Crystal form III is the most structurally different. All other structures superimpose with root mean squared deviation (r.m.s.d.) of <0.7 Å. Circles highlight the most flexible regions. **c**, Comparison between the structures of hNOTUM (form V) and dNotum (form I). The circle highlights the lack of a cysteine bridge in dNotum.



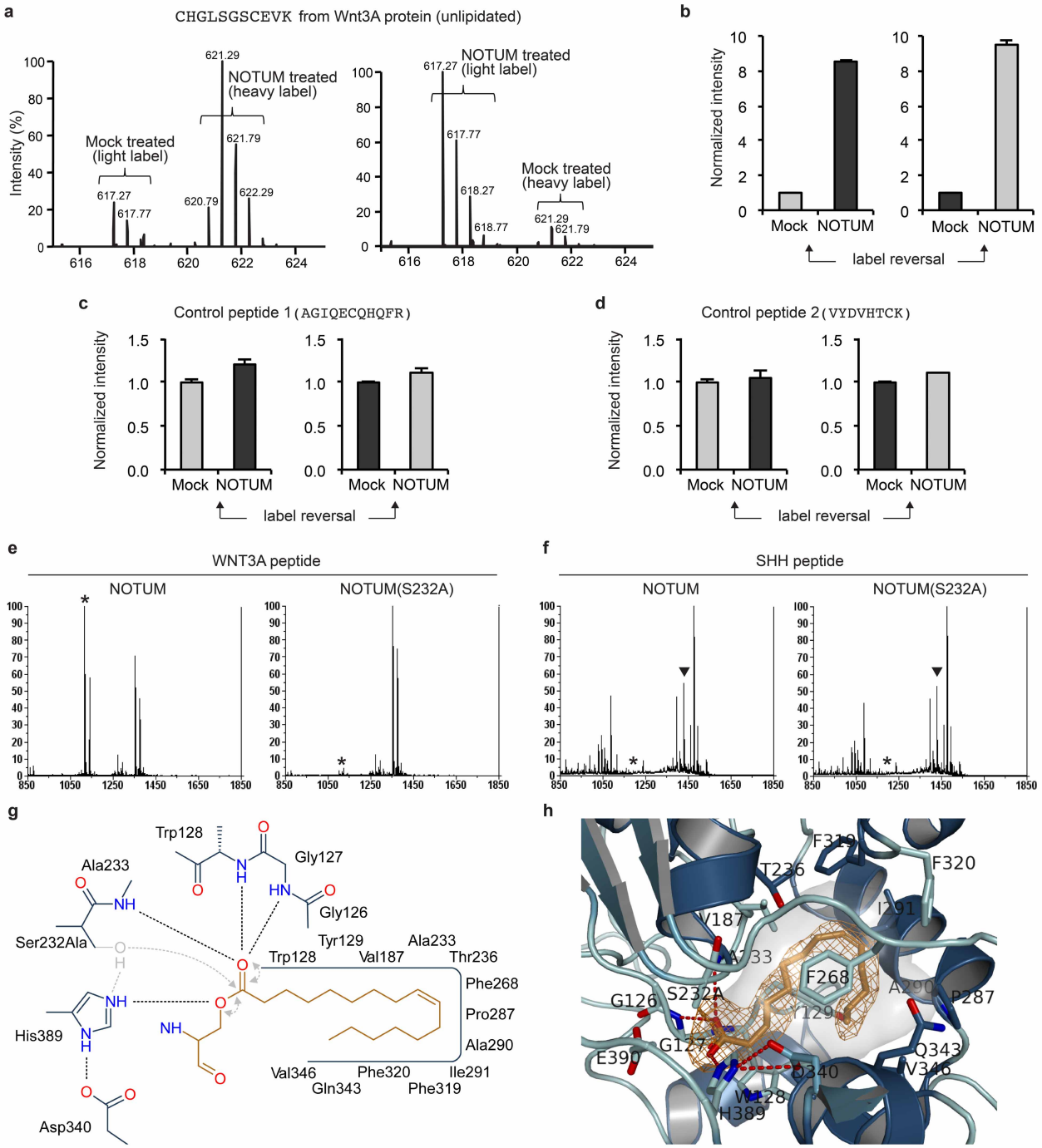
Extended Data Figure 6 | Structural and biophysical analysis of heparin binding. **a**, Heparin affinity chromatography of wild-type hNOTUM and selected surface variants. **b**–**e**, Close-up views of additional sulfate binding sites on hNOTUM, crystal form III. Each view is accompanied with SPR heparin affinity data corresponding to each hNOTUM variant.



Extended Data Figure 7 | Relation of Notum to other esterases of the α/β hydrolase family. **a**, Comparison between hNOTUM and human esterase D (hESTD), showing structural relatedness. hNOTUM is also related to hAPT1, a cytosolic esterase used in this study as a positive control for fatty acid esterase activity. In the views shown here, the hNOTUM structure has been rotated by 90° around the *x* axis relative to the structure shown in Fig. 3b. **b**, Rootless phylogenetic tree of animal Notum proteins (red) and plant pectin acetylesterases (PAE, green). Extent of sequence identity to hNOTUM is shown next to species name. Percentages between branches indicate sequence identity between neighbours.



Extended Data Figure 8 | Substrates and inhibitors of hNOTUM. **a**, Inhibition of hNOTUM activity on pNP-butyrate (pNP4) by PMSF (30 min pre-incubation with 2 mM PMSF) as well as by Triton X-100 and CHAPS (0.5%). Presence of 20 mM SOS and 50 mgl^{-1} heparin results in a minor increase of esterase activity. The height of each bar represents activity relative to the mean of four control samples lacking the additives. **b**, Saturable inhibition of hNOTUM by Triton X-100. Triton X-100 inhibits many esterases owing to binding to the acyl binding pocket through its hydrophobic group. **c**, Lack of inhibition of Norrin-mediated β-catenin stabilization by Notum. Recombinant Norrin was pretreated with $hNOTUM_{core}$ at a concentration sufficient to suppress Wnt3A-mediated signalling. **d**, **e**, Saturation kinetics of the action of hNOTUM on pNP-octanoate (pNP8, **d**) and pNP-butyrate (pNP4, **e**). The activity was normalized to the A_{max} calculated for $hNOTUM_{core}$. The activity values for the larger, full length protein were adjusted to compensate for the increased mass. Apparent K_m values in **d** were corrected for the inhibition caused by Triton X-100. **f**, Saturation inhibition kinetics with myristoleic and palmitoleic acid. pNP8 was used at a concentration of 1 mM and 250 µM, respectively.



Extended Data Figure 9 | Additional mass spectrometric analysis of the hNOTUM deacylase activity. **a**, Mass spectra of CHGLSGSCEVK from trypsinized Wnt3A protein mock-treated or treated with $hNOTUM_{core}$. Left-hand graph is the same as that shown in Fig. 5a, while the right-hand side shows the results of a separate experiment performed with the labels reversed. **b**, Duplicate LC–MS peak areas with label reversal. Irrespective of the nature of the label (grey indicates light label, black indicates heavy label), $hNOTUM_{core}$ triggered an increase in peak area of the delipidated Wnt3A tryptic peptide. **c**, **d**, Two control Wnt3A cysteine-containing peptides from the same data set were not affected by $hNOTUM_{core}$. **e**, Activity of $hNOTUM_{core}$ and its Ser232Ala variant on a synthetic disulphide-bonded Wnt3A peptide (CHGLSGSCEVK) palmitoleoylated on the first serine. Both lipidated and unlipidated peptide could be detected by MALDI-TOF. Incubation with $hNOTUM_{core}$, but not its Ser232Ala variant, caused significant delipidation (peak corresponding to delipidated peptide is marked by asterisk). Quantification of triplicate experiments is shown in Fig. 5c. **f**, MALDI-TOF analysis shows that neither $hNOTUM_{core}$ nor its Ser232Ala variant delipidated a synthetic SHH peptide (CGPGRGFGKRR) palmitoylated on its N-terminal cysteine. Quantification of triplicate experiments is shown in Fig. 5d (peak corresponding to lipidated peptide is marked by black triangle). **g**, Two-dimensional active site schematic relating to Fig. 5e. Additional hydrogen bonds and electron pair movements thought to occur during hydrolysis by the wild type protein are shown in grey. **h**, Close-up view on the myristoleate active site complex of $hNOTUM_{core}$ (crystal form I). The experimental omit electron density is contoured at 2σ.

ARTICLE

doi:10.1038/nature14237

Integrase-mediated spacer acquisition during CRISPR–Cas adaptive immunity

James K. Nuñez[1], Amy S. Y. Lee[1,2], Alan Engelman[3] & Jennifer A. Doudna[1,2,4,5,6]

Bacteria and archaea insert spacer sequences acquired from foreign DNAs into CRISPR loci to generate immunological memory. The *Escherichia coli* Cas1–Cas2 complex mediates spacer acquisition *in vivo*, but the molecular mechanism of this process is unknown. Here we show that the purified Cas1–Cas2 complex integrates oligonucleotide DNA substrates into acceptor DNA to yield products similar to those generated by retroviral integrases and transposases. Cas1 is the catalytic subunit and Cas2 substantially increases integration activity. Protospacer DNA with free 3′-OH ends and supercoiled target DNA are required, and integration occurs preferentially at the ends of CRISPR repeats and at sequences adjacent to cruciform structures abutting AT-rich regions, similar to the CRISPR leader sequence. Our results demonstrate the Cas1–Cas2 complex to be the minimal machinery that catalyses spacer DNA acquisition and explain the significance of CRISPR repeats in providing sequence and structural specificity for Cas1–Cas2-mediated adaptive immunity.

Prokaryotic adaptive immunity relies on clustered regularly interspaced short palindromic repeats (CRISPRs) together with CRISPR associated (Cas) proteins to detect and destroy foreign nucleic acids[1,2]. CRISPR loci contain an AT-rich leader sequence followed by repetitive sequence elements flanking ~30 base pair (bp) spacer segments that are transcribed to produce precursor CRISPR RNAs (pre-crRNAs)[3–5]. Spacers are frequently virus- or plasmid-derived, although 'self'-derived spacers from the host chromosome are present in some CRISPR loci[6]. After pre-crRNA processing and assembly with Cas proteins, the resulting surveillance complexes target and cleave foreign nucleic acids bearing sequences complementary to the crRNA spacer sequence[7–12]. How spacer DNA sequences, termed protospacers, are acquired into the host CRISPR locus remains unknown.

Overexpression of Cas1 and Cas2 nucleases, the only Cas proteins found in all CRISPR–Cas systems, leads to the site-specific acquisition of 33 bp protospacers at the leader end of the CRISPR locus in *E. coli*[13–15]. Furthermore, Cas1 and Cas2 function as a complex *in vivo*[16], suggesting that the Cas1–Cas2 complex might possess DNA recombination activity. We reconstituted CRISPR spacer acquisition using purified Cas1 and Cas2 proteins, protospacers and acceptor plasmid DNA, revealing an elegant mechanism in which both the sequence and structural elements of the CRISPR repeats specify spacer integration sites.

Protospacer DNA integration by Cas1–Cas2

To test whether the Cas1–Cas2 complex is sufficient to catalyse DNA recombination *in vitro*, assays were conducted using purified Cas1–Cas2 complex, 33 bp protospacer DNA and an acceptor 'target' plasmid consisting of the pUC19 backbone with an inserted CRISPR locus (pCRISPR) (Fig. 1a). Co-incubation of these reagents converted the supercoiled plasmid into three main products: relaxed and linear plasmid species and a fast-migrating species we term band X (Fig. 1b, c and Extended Data Fig. 1a). Product formation required Cas1, Cas2 and the protospacer DNA (Extended Data Fig. 1b–d), and was consistent with previous divalent metal ion-dependent and sequence-nonspecific *in vitro* activity requirements of Cas1 (refs 17–19) and Cas2 (refs 20–22). Product DNA migration was not affected by treatment with EDTA, EDTA and phenol–chloroform extraction or proteinase K in the presence of EDTA and detergent (Extended Data Fig. 1e), indicating that product DNAs are unlikely to be bound to Cas1 and/or Cas2. Consistent with product DNA resulting from covalent integration of protospacer DNA into the plasmid, the relaxed and linear forms of pCRISPR became radiolabelled in reactions containing ^{32}P-labelled protospacer DNA (Fig. 1d and Extended Data Fig. 2). Although Cas1 alone catalysed a low level of protospacer integration in the presence of Mn^{2+}, the reaction was enhanced substantially by the presence of Cas2 (Extended Data Fig. 2b).

Bacteria expressing Cas1 active-site mutants, but not Cas2 active-site mutants, are incapable of acquiring new spacers *in vivo*, demonstrating the catalytic role of Cas1 during spacer acquisition[13,14,16]. Consistent with these data, Cas1 active site mutants H208A and D221A were defective for protospacer integration *in vitro*, whereas the Cas2(E9Q) active-site mutant supported integration (Fig. 1c, e and Extended Data Fig. 3). The Cas2 C-terminal (Δβ6–β7) deletion mutant, which is defective for complex formation with Cas1 and spacer acquisition *in vivo*, failed to support Cas1-mediated integrase activity (Fig. 1c, e). We conclude that our *in vitro* assay recapitulates the *in vivo* functions of Cas1 and Cas2 during spacer acquisition.

Integration and disintegration products

We tested whether the reaction products of Cas1–Cas2-mediated DNA integration resemble those formed by the strand transfer activity of retroviral integrases and cut-and-paste transposases[23–26]. These enzymes generate two main products *in vitro* corresponding to half-site and full-site integration events (Fig. 2a). We observed similar gel mobility of the slowly migrating DNA product generated by Cas1–Cas2 and Nb.BbvCI nickase-digested pCRISPR, consistent with the slow-migrating relaxed DNA species corresponding to half-site products and/or products resulting from full-site integration of one protospacer molecule (Extended Data Fig. 1a). Digestion with EcoRI, which cuts pCRISPR once, converted the reaction products to linear DNAs (Fig. 2b, compare lane 4 to lane 2, and Fig. 2c). We therefore conclude that both the relaxed and band X DNA products comprise unit-sized pCRISPR circles.

[1]Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley, California 94720, USA. [2]Center for RNA Systems Biology, University of California, Berkeley, Berkeley, California 94720, USA. [3]Department of Cancer Immunology and AIDS, Dana-Farber Cancer Institute and Department of Medicine, Harvard Medical School, Boston, Massachusetts 02115, USA. [4]Howard Hughes Medical Institute, University of California, Berkeley, Berkeley, California 94720, USA. [5]Department of Chemistry, University of California, Berkeley, Berkeley, California 94720, USA. [6]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA.

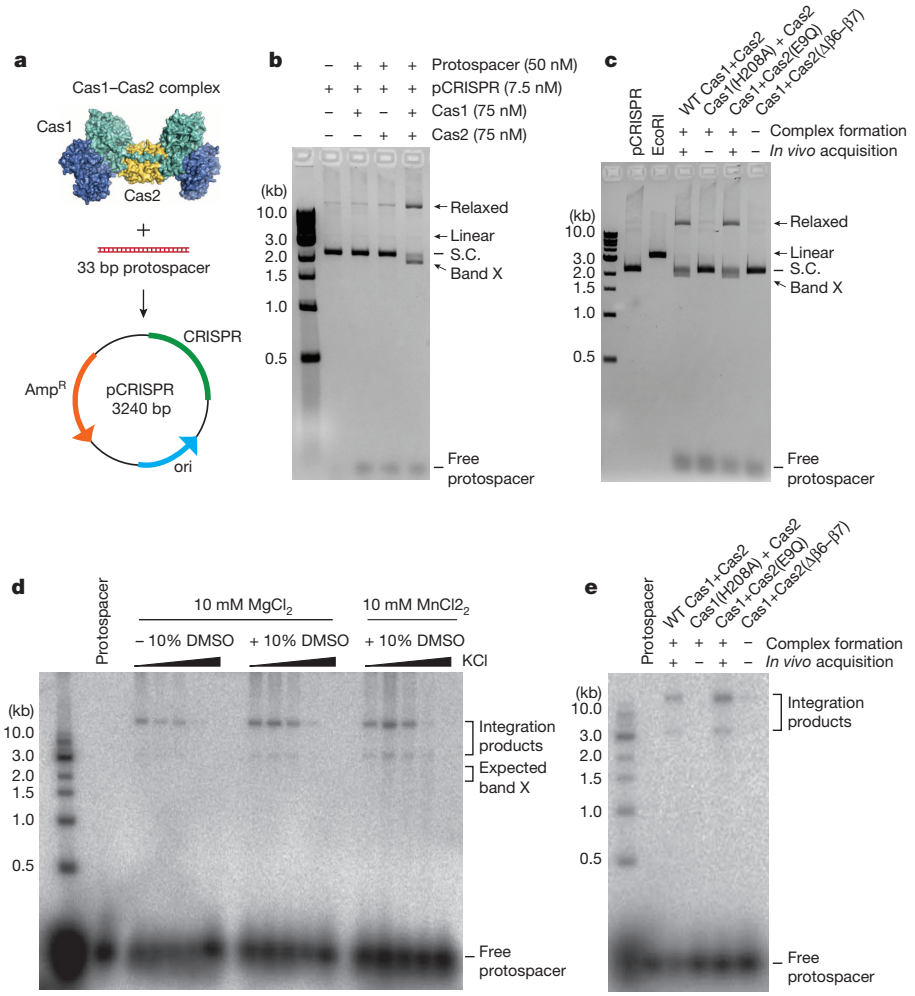


Figure 1 | The Cas1–Cas2 complex integrates protospacers *in vitro*. **a**, Schematic of the *in vitro* integration assay (PDB code 4P6I for Cas1–Cas2). **b**, The presence of Cas1, Cas2 and a protospacer results in the conversion of the supercoiled pCRISPR into relaxed, linear and band X products. **c**, Neither the Cas1(H208A) active site mutant nor and the complex formation-defective Cas2(Δβ6–β7) deletion mutant support the reaction. The Cas2(E9Q) active site mutant is as active as the wild type. **d**, Salt- and metal-dependence of radiolabelled protospacer integration into pCRISPR. **e**, Same as **c** except using radiolabelled protospacers. The data presented in **b**–**e** are representative of at least three replicates.

We observed that band X did not become radiolabelled in reactions conducted with ^{32}P-labelled protospacer DNA. A time-course analysis revealed relaxed DNA product formation within the first minute, followed by accumulation of band X between 10 and 30 min (Fig. 2d). To determine the properties of band X, the purified product was analysed in two different types of agarose gels—one pre-stained with ethidium bromide, similar to the gels presented thus far, and the other stained with ethidium bromide after electrophoresis (post-stained) (Extended Data Fig. 4a). Although band X migrated as a single species in the pre-stained gel, a ladder of species that migrated faster than the relaxed products was observed in the post-stained gel (Fig. 2e, f). These intermediates are reminiscent of plasmid topoisomers[27,28]. The same pre- and post-stained agarose gel analysis was performed on the entire integration reaction, generating similar results to those observed with purified band X (Extended Data Fig. 4b, c). PCR analysis of various segments of pCRISPR using gel-purified band X as the template yielded amplification products indistinguishable from those generated using unreacted supercoiled pCRISPR or relaxed integration products, supporting the conclusion that band X corresponds to pCRISPR topoisomers (Extended Data Fig. 4d).

We wondered whether Band X arose from protospacer excision from half-site integration products to regenerate pCRISPR in different supercoiled states, analogous to the *in vitro* reversal of integration activity of retroviral integrases and transposases (termed disintegration, Fig. 2g)[29,30]. To test this hypothesis, a synthetic Y-structured DNA intermediate that mimics the half-site integration product (Extended Data Fig. 5a, b) was radiolabelled such that the liberated 33 bp protospacer DNA could be detected following disintegration activity. Using this substrate, we observed that Cas1 catalysed disintegration activity either by itself or in the presence of Cas2 (Fig. 2h). Disintegration activity was confirmed by radiolabelling the 20 nucleotide (nt) target DNA strand and monitoring the formation of the joined 40 bp target DNA product (Extended Data Fig. 5c, d). Thus, Cas1–Cas2 integration and disintegration activities are similar to those of retroviral integrases and transposases.

Integration requires 3′-OH protospacer ends

We next investigated the DNA protospacer and target DNA requirements for integration. Single-stranded protospacer DNA failed to support the reaction (Fig. 3a, b). The Cas1–Cas2 complex accommodated various protospacer lengths *in vitro* despite the strict 33 bp requirement for spacer acquisition *in vivo* (Extended Data Fig. 6a), suggesting that protospacer length is pre-determined before integration *in vivo* by an unknown mechanism. The Cas1–Cas2 complex integrated DNA substrates with blunt-ends or with 3′-overhangs up to 5 nt in length (Extended Data Fig. 6b). In contrast to retroviral integrases[31], substrates with 5′-overhangs were non-viable (Extended Data Fig. 6b).

Retroviral integration and transposition reactions proceed via nucleophilic attack of DNA 3′-OH groups at target DNA phosphodiester bonds[31,32]. We found that phosphorylation of both 3′-ends of the protospacer ablated integration, whereas phosphorylation of only one 3′ end strongly limited integration (Fig. 3a, b). By analogy to known integrase enzyme mechanisms, DNA integration could proceed by Cas1-catalysed direct nucleophilic attack of the substrate 3′-OH on the target DNA, or by formation of a Cas1–DNA intermediate, as occurs in the serine and tyrosine families of recombinases[33]. Four tyrosine residues in the vicinity of the Cas1 active site[17–19] could be involved in forming such a covalent intermediate (Extended Data Fig. 7a, b). Purified Cas1 mutant proteins in which each tyrosine was individually changed to

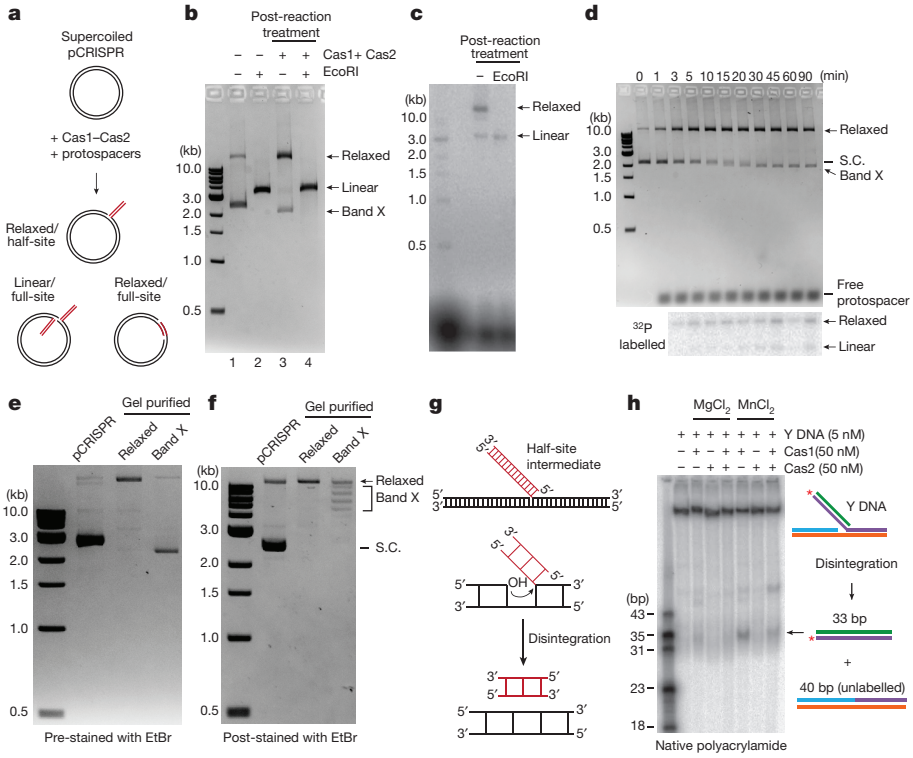


Figure 2 | Half-site, full-site integration and pCRISPR topoisomer products. **a**, Schematic of half-site and full-site integration products. **b**, Linearization of the integration products (lane 4). Lane 3 is the untreated reaction products. **c**, Linearization of integration products from radiolabelled protospacer reactions. **d**, The time course reveals the initial formation of relaxed products, followed by band X. The inset reveals the products detected using ^{32}P-labelled protospacers. **e**, **f**, Analysis of gel-purified relaxed and band X on agarose gels pre-stained with ethidium bromide (**e**) or post-stained after electrophoresis (**f**). **g**, Schematic of the disintegration reaction. **h**, Native polyacrylamide gel analysis of the disintegration reaction. The data presented in **b**–**f**, **h** are representative of at least three replicates.

alanine supported protospacer integration *in vitro* at levels comparable to wild-type Cas1–Cas2 (Extended Data Fig. 7c). Thus, the integration reaction likely proceeds via direct nucleophilic attack of protospacer 3′-OH ends onto the target DNA phosphodiester bonds, a mechanism previously hypothesized to occur *in vivo*[34].

Supercoiled DNA and CRISPR locus requirements

Cas1 and Cas2 overexpression leads to site-selective spacer acquisition proximal to the leader end of the CRISPR locus, a result consistent with observations in native populations of CRISPR-containing bacteria[13–15,35]. To determine what drives such site-specific integration, we first tested various forms of the pCRISPR plasmid to determine target DNA requirements. Integration requires target DNA supercoiling, as neither relaxed nor linear pCRISPR, nor the isolated 1 kb CRISPR locus, supported integration (Fig. 3c and Extended Data Fig. 6c, d).

As a control, we tested supercoiled pUC19 DNA, the parental plasmid of pCRISPR that lacks a CRISPR locus, and were surprised to observe integration products upon incubation with Cas1 and Cas2 in the presence of protospacer DNA (Fig. 3c and Extended Data Fig. 6e). This finding raised two possibilities: either *in vitro* spacer integration is non-specific with respect to target DNA sequence or structures and/or sequence(s) favouring integration are present in the pUC19 plasmid. To determine if integration preferentially occurred at the CRISPR locus of pCRISPR, products of radiolabelled reactions were double-digested to separate the CRISPR locus (960 bp) from the pUC19 plasmid backbone (~2.27 kb). Suggestive of CRISPR-specific integration, the ^{32}P-radiolabel migrated solely with the CRISPR locus fragment (Fig. 3d). The same result was observed when the experiment was conducted using a target plasmid containing the CRISPR locus and a different backbone sequence (pACYC) (Fig. 3e).

CRISPR repeats provide specificity

To determine the exact sites of protospacer integration in these reactions, we performed high-throughput sequencing of reaction products

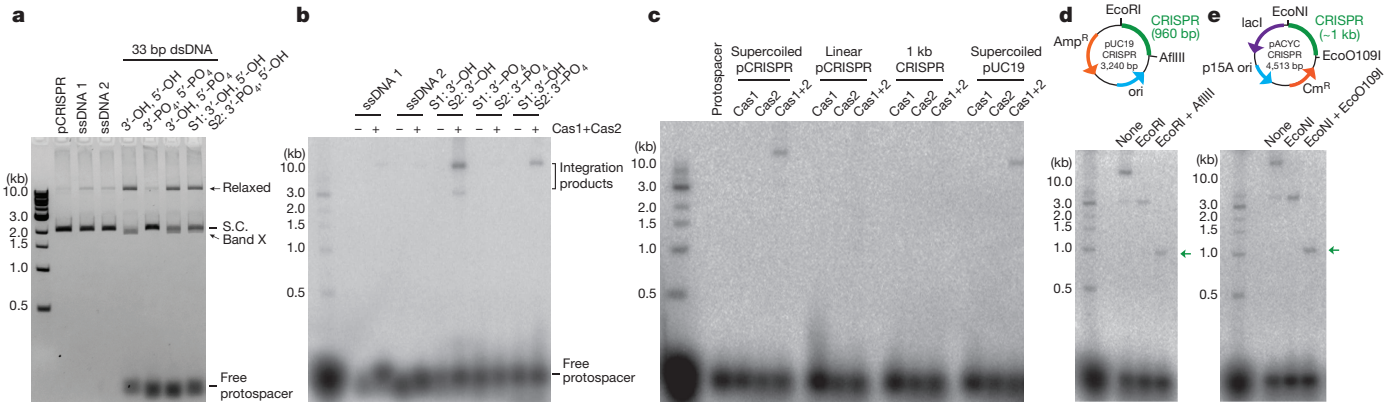


Figure 3 | Integration requires 3′-OH protospacer ends and supercoiled target DNA. **a**, **b**, Integration assays using single-stranded DNAs and either –OH or –PO_4 at the 3′ or 5′ ends of unlabelled (**a**) or radiolabelled (**b**) protospacers. S1 corresponds to one strand of the protospacer and S2 corresponds to the complementary strand. **c**, Comparison of protospacer integration into different DNA targets. **d**, **e**, Restriction enzyme digestion of pCRISPR, either in a pUC19 (**d**) or pACYC backbone (**e**), after the integration assay detects integration into the CRISPR fragment (green arrows). The data presented in **a**–**e** are representative of at least three replicates.

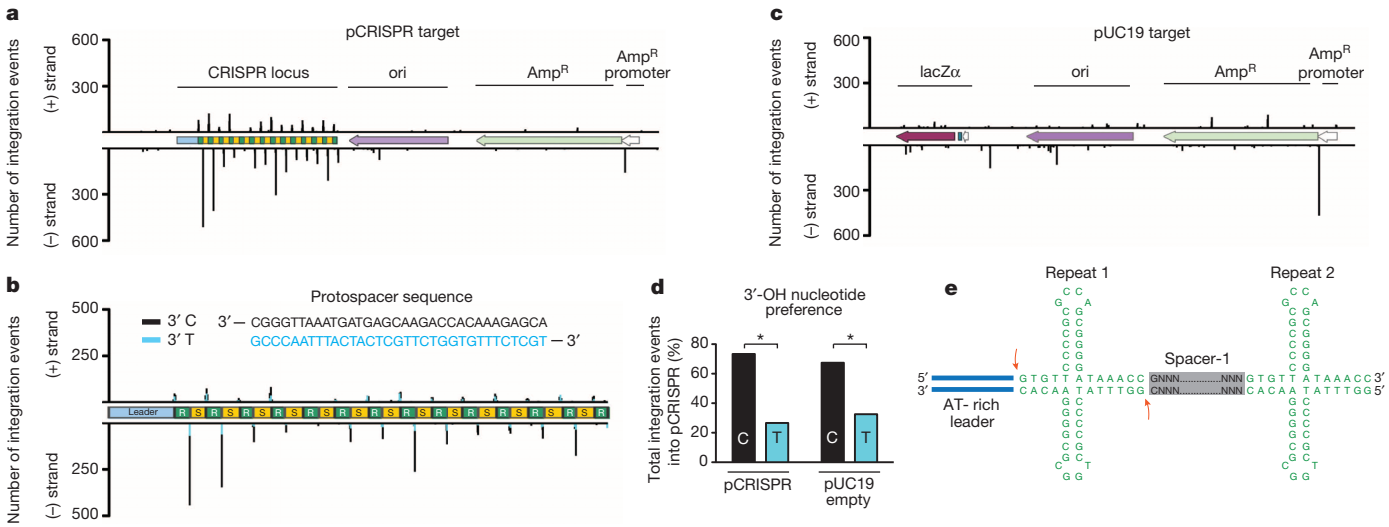


Figure 4 | Protospacers are specifically integrated into the CRISPR locus. **a**, Integration sites along pCRISPR. **b**, Magnified view of the integration sites along the ~1 kb CRISPR locus. The cyan peaks represent positions where the 3′ T of the protospacer DNA was integrated whereas the black peaks represent the C 3′-OH integration events. The protospacer sequence is depicted above the plot. **c**, Integration sites along pUC19. **d**, Comparison of C 3′-OH or T 3′-OH selection in the total reads from pCRISPR and pUC19 targets ($n = 7{,}866$ reads for pCRISPR and $n = 5{,}524$ reads for pUC19, chi-square test, $^{*}P < 0.0001$). **e**, Schematic of DNA cruciform formation of the repeat sequences. The orange arrows depict the cleavage sites.

that resulted from using either pCRISPR or the parental pUC19 vector as the target of integration (Extended Data Fig. 8a). Of the 7,866 protospacer-pCRISPR junctions retrieved, ~71% mapped to the CRISPR locus (Fig. 4a and Extended Data Fig. 8b). Protospacer insertion occurred at the borders of each repeat, with the most preferred site at the first repeat adjacent to the leader (Fig. 4b). The minus strand of each repeat (the bottom strand in Fig. 4a, b that runs 5′ to 3′ towards the leader sequence) is also highly preferred, highlighting the role of CRISPR repeats in providing sequence specificity for the Cas1–Cas2 complex (Fig. 4b). Sequence alignment of the integration sites revealed strong preference for sequences resembling the CRISPR repeat on both strands of pCRISPR, further supporting the selection of CRISPR repeat borders by the Cas1–Cas2 complex (Extended Data Fig. 8d–f).

The most frequent integration site in the pUC19 control plasmid mapped to the *amp* resistance gene adjacent to the AT-rich promoter sequence (~8.8% of 5,524 total retrieved junctions, Fig. 4c and Extended Data Fig. 8c). An inverted repeat sequence with a propensity to form a DNA cruciform[36] occurs 9 nt adjacent to this integration site (plus strand sequence: 5′-TTCAATATTATTGAA-3′), suggesting that potential DNA cruciform formation adjacent to AT-rich sequences is important for protospacer integration. Sequence analysis of pUC19 target sites revealed the propensity for a G nucleotide to occur at the −2 and +1 positions of the protospacer insertion site, similar to the preferred pCRISPR sites (Extended Data Fig. 8g, h). These observations imply that in addition to sequence, pCRISPR repeat selectivity stems from the unique structural features of these sites, such as their ability to form cruciforms (Fig. 4a, b, e).

In *E. coli*, newly acquired spacers harbour a 5′ G as the first nucleotide flanking the leader-proximal end of the repeats, which originates from the last nucleotide of the AAG protospacer-adjacent motif (PAM) from foreign DNA[13–15,37–39]. Such positional specificity is critical for crRNA-guided interference, as a mutation in this position of the corresponding crRNA disrupts PAM binding and subsequent target destruction[40–42]. We found that ~73% of all integration events into pCRISPR used the 3′ C end instead of the 3′ T end of protospacer DNA during integration (see Fig. 4b for protospacer sequence), and there was a strong preference for this nucleotide to attack the minus strand of the repeat sequence (Fig. 4b, d, e). A similar nucleotide bias was observed in the pUC19 target plasmid sequence data (Fig. 4d). This preference positions the G at the 5′ end of the protospacer substrate as the first nucleotide of the newly integrated spacer in the CRISPR locus (Fig. 5).

When we used protospacer DNAs lacking a 3′ C or bearing 3′ C on both ends, the preference for integration into the minus strand of the CRISPR locus was significantly decreased (Extended Data Fig. 9). Thus, the Cas1–Cas2 complex plays a critical role in correctly orienting the C 3′-OH end of protospacer DNA substrates for incorporation within the CRISPR locus.

Mechanism of protospacer integration

The results presented here explain the mechanistic basis for foreign DNA acquisition during CRISPR–Cas adaptive immunity (Fig. 5). The Cas1–Cas2 complex catalyses integration of protospacers at the leader-end of

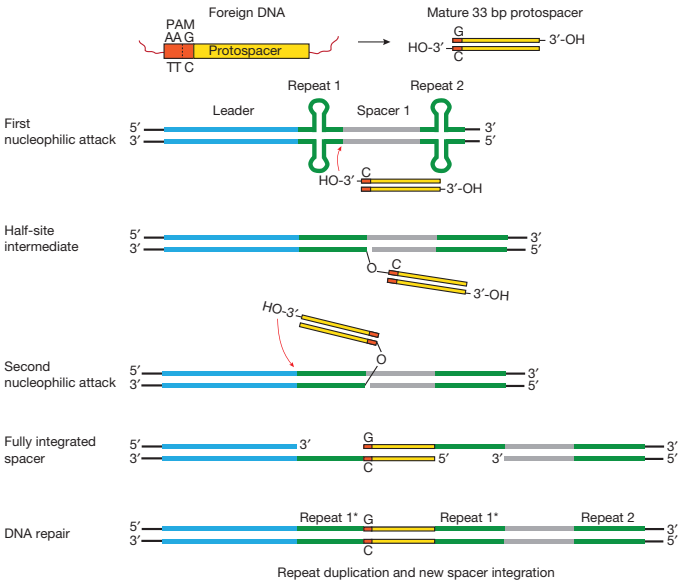


Figure 5 | Model of protospacer integration during CRISPR–Cas adaptive immunity. The first nucleophilic attack occurs on the minus strand of the first repeat, distal to the leader, by the C 3′-OH end of the protospacer. After half-site intermediate formation, the second integration event occurs on the opposite strand at the leader-repeat border. The resulting single-stranded DNA gaps are repaired by yet uncharacterized mechanisms and the protospacer is fully integrated with the G as the first nucleotide at its 5′ end. The asterisk denotes the duplication of the first repeat, as previously observed *in vivo*[13–15].

the CRISPR locus and also selects the terminal C 3′-OH as the attacking nucleophile, resulting in the 5′ G on the opposite strand of the protospacer becoming the first nucleotide of the newly integrated spacer. This orientation bias, previously observed *in vivo*[39], is a key step during immunity for productive downstream foreign DNA targeting by the Cascade complex and Cas3 effector nuclease (Extended Data Fig. 10). Interestingly, the presence of the complete AAG PAM in the protospacer is not required for *in vitro* integration, suggesting that a highly specific selection or processing step occurs *in vivo* to exclude the AA nucleotides from the mature protospacer before integration.

We propose a two-step integration mechanism in which the C 3′-OH first attacks the minus strand of the CRISPR repeat to produce a half-site intermediate (Fig. 5). The 3′-OH on the opposite strand of the integrating DNA then attacks the target DNA 28 bp away on the opposite side of the repeat on the plus strand, leading to full integration of the protospacer (Fig. 5). Our *in vitro* system predominantly traps the first step of this two-step integration mechanism, suggesting that the second nucleophilic attack is greatly accelerated *in vivo* in the presence of cellular factors. This model is consistent with spacer integration intermediates that are observed *in vivo*, in which protospacers are integrated such that staggered cleavage at each end of the repeat generates single-stranded gaps that ensure repeat duplication[34]. The *in vivo* conditions could also promote the high specificity of integration to occur solely downstream of the first repeat of the CRISPR locus in *E. coli*, instead of at every repeat, as observed in our *in vitro* assay.

CRISPR spacer integration shares mechanistic similarities with retroviral integration and DNA transposition, where the integrase/transposase enzyme uses donor DNA 3′-OH ends to make a staggered cut at the DNA target site, which concurrently joins the donor DNA to target DNA 5′-phosphates[31,32]. Completion of the integration reaction requires a DNA polymerase to fill in sequence gaps and a DNA ligase to seal the phosphodiester backbone[43]. Similar polymerase and ligase functions are required to complete CRISPR spacer acquisition *in vivo*, although the specific enzymes involved have not yet been identified. Despite these similarities, we note that the Cas1 active site does not harbour the RNase H fold that defines the retroviral integrase enzyme superfamily[44]. This structural difference could explain the unexpected production of different topoisomers of pCRISPR (band X) *in vitro*, although the physiological significance of band X production remains unclear.

Our results highlight the fundamental role of repeat sequences at multiple stages of CRISPR–Cas adaptive immunity. In addition to creating structures within nascent CRISPR transcripts that ensure correct RNA processing during crRNA maturation[45], the repeats operate at the DNA level to recruit the Cas1–Cas2 complex for sequence- and structure-specific protospacer integration. We envision that this recruitment involves transient DNA cruciform formation within the CRISPR inverted repeats that occurs as a function of target DNA supercoiling[46]. The observation that a preferred non-CRISPR site of Cas1–Cas2-mediated DNA integration is proximal to an inverted repeat adjacent to an AT-rich sequence suggests the fascinating possibility that CRISPR loci arise in naive genomes through integration events that become self-propagating through creation of repetitive sequences with properties that ensure continual recognition and activity by the Cas1–Cas2 integration machinery.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 6 November 2014; accepted 15 January 2015.
Published online 18 February 2015.

1. Barrangou, R. *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315,** 1709–1712 (2007).
2. van der Oost, J., Westra, E. R., Jackson, R. N. & Wiedenheft, B. Unravelling the structural and mechanistic basis of CRISPR–Cas systems. *Nature Rev. Microbiol.* **12,** 479–492 (2014).
3. Mojica, F. J., Diez-Villasenor, C., Garcia-Martinez, J. & Soria, E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J. Mol. Evol.* **60,** 174–182 (2005).
4. Bolotin, A., Quinquis, B., Sorokin, A. & Ehrlich, S. D. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151,** 2551–2561 (2005).
5. Pourcel, C., Salvignol, G. & Vergnaud, G. CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology* **151,** 653–663 (2005).
6. Stern, A., Keren, L., Wurtzel, O., Amitai, G. & Sorek, R. Self-targeting by CRISPR: gene regulation or autoimmunity? *Trends in Genet.* **26,** 335–340 (2010).
7. Carte, J., Wang, R., Li, H., Terns, R. M. & Terns, M. P. Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. *Genes Dev.* **22,** 3489–3496 (2008).
8. Haurwitz, R. E., Jinek, M., Wiedenheft, B., Zhou, K. & Doudna, J. A. Sequence- and structure-specific RNA processing by a CRISPR endonuclease. *Science* **329,** 1355–1358 (2010).
9. Deltcheva, E. *et al.* CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471,** 602–607 (2011).
10. Brouns, S. J. *et al.* Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321,** 960–964 (2008).
11. Garneau, J. E. *et al.* The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468,** 67–71 (2010).
12. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337,** 816–821 (2012).
13. Yosef, I., Goren, M. G. & Qimron, U. Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. *Nucleic Acids Res.* **40,** 5569–5576 (2012).
14. Datsenko, K. A. *et al.* Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. *Nature Commun.* **3,** 945 (2012).
15. Swarts, D. C., Mosterd, C., van Passel, M. W. & Brouns, S. J. CRISPR interference directs strand specific spacer acquisition. *PLoS ONE* **7,** e35888 (2012).
16. Nuñez, J. K. *et al.* Cas1–Cas2 complex formation mediates spacer acquisition during CRISPR–Cas adaptive immunity. *Nature Struct. Mol. Biol.* **21,** 528–534 (2014).
17. Wiedenheft, B. *et al.* Structural basis for DNase activity of a conserved protein implicated in CRISPR-mediated genome defense. *Structure* **17,** 904–912 (2009).
18. Babu, M. *et al.* A dual function of the CRISPR–Cas system in bacterial antivirus immunity and DNA repair. *Mol. Microbiol.* **79,** 484–502 (2011).
19. Kim, T. Y., Shin, M., Huynh Thi Yen, L. & Kim, J. S. Crystal structure of Cas1 from Archaeoglobus fulgidus and characterization of its nucleolytic activity. *Biochem. Biophys. Res. Commun.* **441,** 720–725 (2013).
20. Beloglazova, N. *et al.* A novel family of sequence-specific endoribonucleases associated with the clustered regularly interspaced short palindromic repeats. *J. Biol. Chem.* **283,** 20361–20371 (2008).
21. Samai, P., Smith, P. & Shuman, S. Structure of a CRISPR-associated protein Cas2 from *Desulfovibrio vulgaris*. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **66,** 1552–1556 (2010).
22. Nam, K. H. *et al.* Double-stranded endonuclease activity in *Bacillus halodurans* clustered regularly interspaced short palindromic repeats (CRISPR)-associated Cas2 protein. *J. Biol. Chem.* **287,** 35943–35952 (2012).
23. Li, M. & Craigie, R. Processing of viral DNA ends channels the HIV-1 integration reaction to concerted integration. *J. Biol. Chem.* **280,** 29334–29339 (2005).
24. Cherepanov, P. LEDGF/p75 interacts with divergent lentiviral integrases and modulates their enzymatic activity *in vitro*. *Nucleic Acids Res.* **35,** 113–124 (2007).
25. Hare, S. *et al.* A novel co-crystal structure affords the design of gain-of-function lentiviral integrase mutants in the presence of modified PSIP1/LEDGF/p75. *PLoS Pathog.* **5,** e1000259 (2009).
26. Yang, J. Y., Jayaram, M. & Harshey, R. M. Positional information within the Mu transposase tetramer: catalytic contributions of individual monomers. *Cell* **85,** 447–455 (1996).
27. Dinardo, S., Voelkel, K. A., Sternglanz, R., Reynolds, A. E. & Wright, A. *Escherichia coli* DNA topoisomerase I mutants have compensatory mutations in DNA gyrase genes. *Cell* **31,** 43–51 (1982).
28. Pruss, G. J., Manes, S. H. & Drlica, K. *Escherichia coli* DNA topoisomerase I mutants: increased supercoiling is corrected by mutations near gyrase genes. *Cell* **31,** 35–42 (1982).
29. Chow, S. A., Vincent, K. A., Ellison, V. & Brown, P. O. Reversal of integration and DNA splicing mediated by integrase of human immunodeficiency virus. *Science* **255,** 723–726 (1992).
30. Au, T. K., Pathania, S. & Harshey, R. M. True reversal of Mu integration. *EMBO J.* **23,** 3408–3420 (2004).
31. Engelman, A., Mizuuchi, K. & Craigie, R. HIV-1 DNA integration: mechanism of viral DNA cleavage and DNA strand transfer. *Cell* **67,** 1211–1221 (1991).
32. Mizuuchi, K. & Adzuma, K. Inversion of the phosphate chirality at the target site of Mu DNA strand transfer: evidence for a one-step transesterification mechanism. *Cell* **66,** 129–140 (1991).
33. Curcio, M. J. & Derbyshire, K. M. The outs and ins of transposition: from mu to kangaroo. *Nature Rev. Mol. Cell Biol.* **4,** 865–877 (2003).
34. Arslan, Z., Hermanns, V., Wurm, R., Wagner, R. & Pul, U. Detection and characterization of spacer integration intermediates in type I-E CRISPR-Cas system. *Nucleic Acids Res.* **42,** 7884–7893 (2014).
35. Tyson, G. W. & Banfield, J. F. Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. *Environ. Microbiol.* **10,** 200–207 (2008).
36. Sheflin, L. G. & Kowalski, D. Altered DNA conformations detected by mung bean nuclease occur in promoter and terminator regions of supercoiled pBR322 DNA. *Nucleic Acids Res.* **13,** 6137–6154 (1985).

37. Goren, M. G., Yosef, I., Auster, O. & Qimron, U. Experimental definition of a clustered regularly interspaced short palindromic duplicon in *Escherichia coli*. *J. Mol. Biol.* **423,** 14–16 (2012).
38. Savitskaya, E., Semenova, E., Dedkov, V., Metlitskaya, A. & Severinov, K. High-throughput analysis of type I-E CRISPR/Cas spacer acquisition in *E. coli*. *RNA Biol.* **10,** 716–725 (2013).
39. Shmakov, S. *et al.* Pervasive generation of oppositely oriented spacers during CRISPR adaptation. *Nucleic Acids Res.* **42,** 5907–5916 (2014).
40. Deveau, H. *et al.* Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. *J. Bacteriol.* **190,** 1390–1400 (2008).
41. Semenova, E. *et al.* Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. *Proc. Natl Acad. Sci. USA* **108,** 10098–10103 (2011).
42. Westra, E. R. *et al.* Type I-E CRISPR-cas systems discriminate target from non-target DNA through base pairing-independent PAM recognition. *PLoS Genet.* **9,** e1003742 (2013).
43. Craigie, R. & Bushman, F. D. HIV DNA integration. *Cold Spring Harbor Perspect. Med.* **2,** a006890 (2012).
44. Nowotny, M. Retroviral integrase superfamily: the structural perspective. *EMBO Rep.* **10,** 144–151 (2009).
45. Hochstrasser, M. L. & Doudna, J. A. Cutting it close: CRISPR-associated endoribonuclease structure and function. *Trends Biochem. Sci.* **40,** 58–66 (2015).
46. Paleček, E. Local supercoil-stabilized DNA structures. *Crit. Rev. Biochem. Mol. Biol.* **26,** 151–226 (1991).

Acknowledgements We are grateful to M. Chung, P. J. Kranzusch and A.V. Wright for technical assistance and members of the Doudna laboratory and J. Cate for discussions. This project was funded by US National Science Foundation grant no. 1244557 to J.A.D. and by NIH grant AI070042 to A.E. This work used the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by NIH S10 Instrumentation Grants S10RR029668 and S10RR027303. J.K.N. is supported by a US National Science Foundation Graduate Research Fellowship and a UC Berkeley Chancellor's Graduate Fellowship. A.S.Y.L. is supported as an American Cancer Society Postdoctoral Fellow (PF-14-108-01-RMC). J.A.D. is an Investigator of the Howard Hughes Medical Institute and a member of the Center for RNA Systems Biology.

Author Contributions J.K.N. performed the biochemical experiments. A.S.Y.L. processed and analysed the high-throughput sequencing data. J.K.N., A.S.Y.L., A.E. and J.A.D. designed the study, analysed the data and wrote the manuscript.

Author Information Sequencing data are deposited in Gene Expression Omnibus under accession number GSE64552. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.A.D. (doudna@berkeley.edu).

METHODS

Cas1, Cas2 and DNA preparation. The *cas1* and *cas2* genes from *E. coli* K12 (MG1655) were cloned into expression vectors and the proteins were separately purified as previously described[16]. The proteins were stored in 100 mM KCl, 20 mM HEPES-NaOH, 5% glycerol and 1 mM TCEP at −80 °C before use. Single-stranded DNAs were synthesized (Integrated DNA Technologies). Double-stranded DNA protospacers were annealed in 20 mM HEPES-NaOH, pH 7.5, 25 mM KCl, 10 mM $MgCl_2$ or $MnCl_2$, 1 mM DTT, 10% DMSO by heating at 95 °C for 3 min and slow cooling to room temperature. The sequence of the 33 bp protospacer used in this study was shown to be the most acquired *in vivo* in *E. coli* K12 after M13 bacteriophage infection[14]: strand 1 (5′-GCCCAATTTACTACTCGTTCTGGTGTTTCT CGT-3′) and strand 2 (5′-ACGAGAAACACCAGAACGAGTAGTAAATTGG GC-3′). The pCRISPR target plasmid was constructed by PCR amplifying the *E. coli* BL21-AI genomic CRISPR locus and cloning the fragment into pUC19 using the following primers with the underlines indicating the respective restriction sites used: forward/EcoRI: 5′-ACGTCGAATTCTACCTTTTTAATCAATGG-3′ and reverse/AflIII: 5′-ACGTCACATGTGGTTATATGGTGGTTTATCC-3′. The pACYC CRISPR plasmid was constructed by cloning the CRISPR fragment into a pACYCDuet-1 vector using the EcoNI and AvrII restriction sites.

***In vitro* integration assays.** The integration reactions were performed in 20 mM HEPES-NaOH, pH 7.5, 25 mM KCl, 10 mM $MgCl_2$ or $MnCl_2$, 1 mM DTT and 10% DMSO. There was little difference when DMSO was omitted from the reaction (Fig. 1d), in contrast to its *in vitro* integration enhancement with HIV-1 integrase[47]. All of the reactions were conducted with $MgCl_2$ unless otherwise noted. For reactions with the Cas1–Cas2 complex, separately purified Cas1 and Cas2 were pre-incubated for 20–30 min at 4 °C to allow complex formation. The protospacer DNAs were incubated with the protein(s) for 10–15 min at 4 °C, followed by the addition of the target pCRISPR or pUC19 plasmid DNA. The reactions were conducted at 37 °C for 1 h and quenched with DNA loading buffer containing a final concentration of 50 mM EDTA. The products were analysed on 1.5% agarose gels pre-stained with ethidium bromide. All of the reactions, except those shown in Fig. 1 and Extended Data Fig. 1a, c–e, were conducted with 75 nM protein, 200 nM protospacers and 7.5 nM pCRISPR to clearly visualize band X from pCRISPR. Reactions in Fig. 1 and Extended Data Fig. 1a, c, e were performed with 50 nM protospacers. Each integration and disintegration assay was performed a minimum of three times.

Radiolabelled protospacer integration assays. Pre-annealed double-stranded protospacer DNA substrates were 5′-radiolabelled using [γ-^{32}P]-ATP (PerkinElmer) and T4 polynucleotide kinase (New England Biolabs). Protospacers with 3′-PO_4 ends were 5′-radiolabelled using T4 polynucleotide kinase with 3′ phosphatase minus activity (New England Biolabs). The reactions were carried out in the same buffer as above. Unless otherwise noted, 200 nM of Cas1–Cas2 was first incubated with 20 nM protospacers at 4 °C for 10–15 min, followed by the addition of 200 ng (~5 nM) of pCRISPR. The reactions were conducted at 37 °C for 1 h and quenched with 25 mM EDTA and 0.4% SDS. The DNA samples were deproteinized with 30 μg of proteinase K for 1 h at 37 °C and ethanol precipitated. The reactions were analysed on 1.5% agarose gels. After electrophoresis, the gels were dried onto positively charged nylon transfer membrane (GE Healthcare) and imaged using Phosphor Screens (GE Healthcare). The restriction enzyme digest experiments were performed by first conducting the integration reaction, followed by addition of the respective enzymes (New England Biolabs), which were allowed to digest for an additional 1 h at 37 °C.

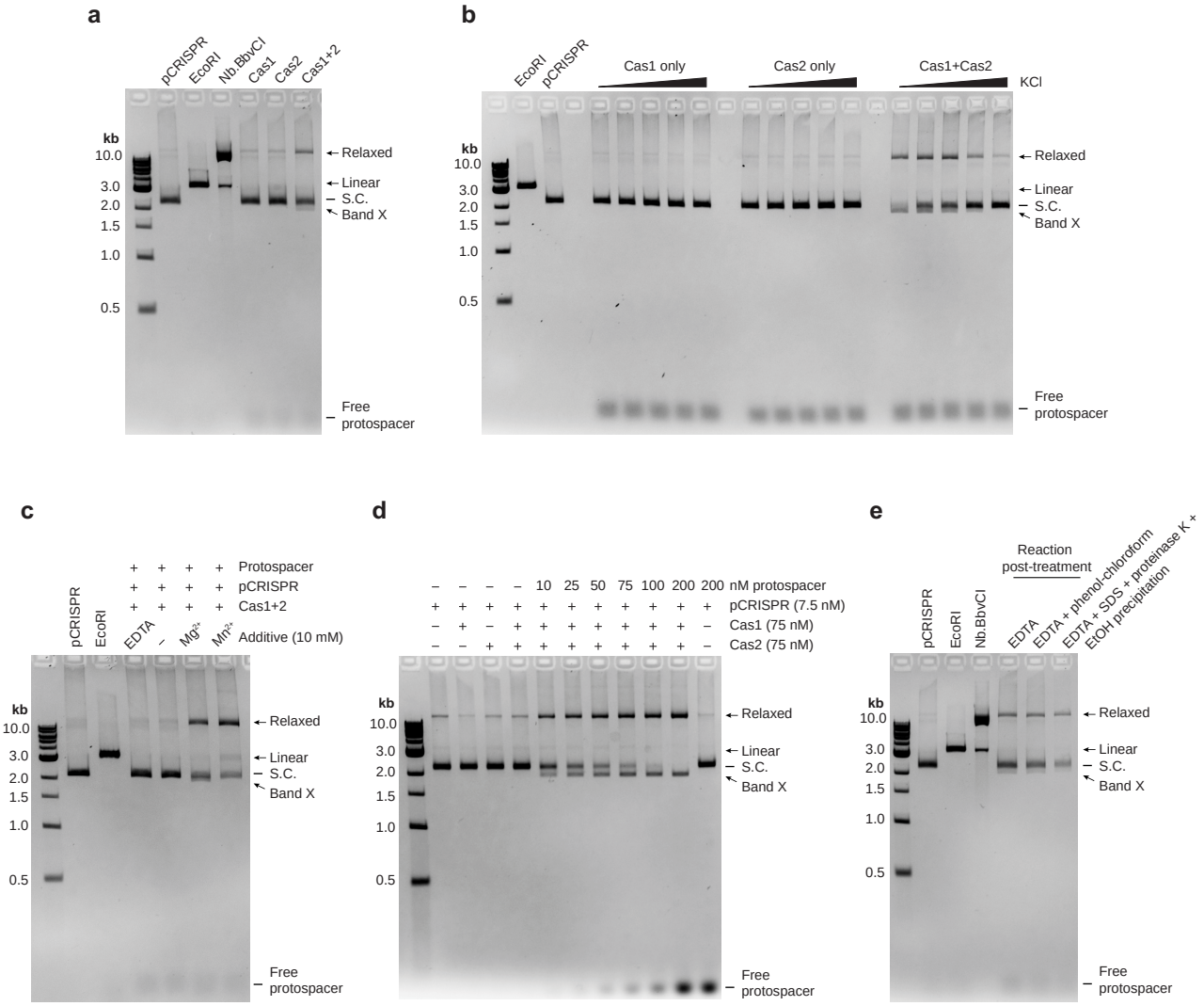
Disintegration assays. The four single-stranded DNA substrates were annealed to form the Y DNA in a stepwise manner: 95 °C for 3 min, 65 °C for 20 min, 50 °C for 20 min, and gradual cooling to room temperature. The annealing reactions were analysed on a 15% native polyacrylamide gel to confirm the formation of the Y DNA (Extended Data Fig. 5b). The disintegration assay was performed in the integration reaction buffer with 50 nM protein and 5 nM Y DNA at 37 °C for 1 h. For native polyacrylamide gel analysis, the reaction was quenched with DNA loading buffer with 50 mM EDTA and analysed on 15% polyacrylamide gels. For denaturing polyacrylamide gel analysis, the reactions were quenched with formamide buffer and heated at 95 °C before loading on 15% 8M urea-polyacrylamide gels. The sequences of the four strands are as follows: A (5′-GGCCCCAGTGCTGCA ATGAT-3′); B (5′-GTGAGCGTGGGTCTCGCGGTATCATTGCAGCACTGG GGCC-3′); C (5′-GCCCAATTTACTACTCGTTCTGGTGTTTCTCGTACCGC GAGACCCACGCTCAC-3′); and D (5′-ACGAGAAACACCAGAACGAGTAG TAAATTGGGC-3′).

High-throughput sequencing. The integration reaction was performed with 75 nM Cas1–Cas2, 200 nM protospacer and 7.5 nM pCRISPR or pUC19 in 20 mM HEPES, pH 7.5, 25 mM KCl, 10 mM $MgCl_2$, 10% DMSO and 1 mM DTT. The DNAs were isolated by phenol–chloroform extraction and ethanol precipitation. The excess protospacers were removed using 100K MWCO Amicon Ultra-0.5 ml centrifugal filters. The resulting integration products were digested into smaller DNA fragments using dsDNA Fragmentase (New England Biolabs) for 75 min at 37 °C and quenched at 65 °C for 15 min. Fragments were end repaired using T4 DNA Polymerase (NEB), Klenow (NEB) and T4 PNK (NEB) and A-tailed with Klenow exo (3′ to 5′ exo minus) (NEB). Adapters were ligated onto fragments using T4 DNA ligase (NEB) and cDNA libraries were amplified by PCR using Phusion (NEB). Libraries were sequenced on an Illumina HiSeq2500 on rapid run mode. The oligonucleotides used are: universal adaptor: 5′-AATGATACGGCGACCACCGA GATCTACACTCTTTCCCTACACGA CGCTCTTCCGATC*T-3′ (*phosphorothioate bond); indexed adaptor: 5′-/5Phos/GATCGGAAGAGCACACGTCTG AACTCCAGTCAC-index-ATCTCGTATGCCGTC TTCTGCTTG-3′); PCR primers: 5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACAC GA-3′, 5′-CAAGCAGAAGACGGCATACGAGAT-3′.

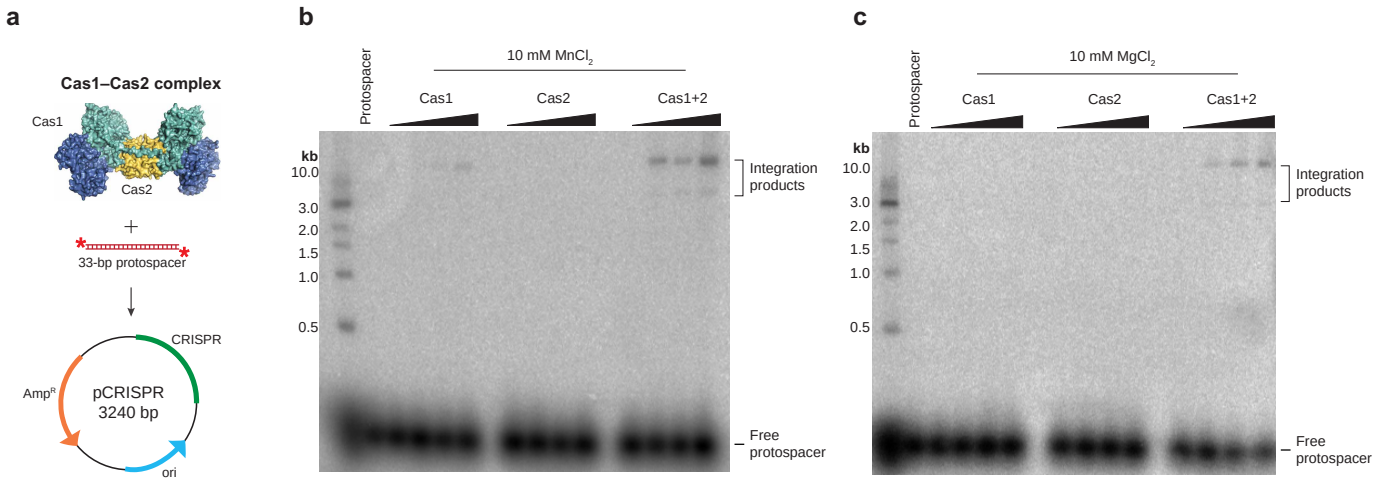
Computational analysis. For preprocessing, 3′ adapters were removed from raw Illumina reads using Cutadapt (http://code.google.com/p/cutadapt/), discarding reads shorter than 15 nt. Reads containing integrated protospacer were selected using Cutadapt, requiring the presence of at least 10 nt of protospacer sequence with no errors. After creating Bowtie[48] indexes from fasta files of the pUC19 empty and pCRISPR plasmid sequences, these reads were mapped to the respective plasmids using Bowtie, allowing up to 2 mismatches and requiring unique mapping. Sequence motif analysis depicted in Extended Data Fig. 8 were generated using WebLogo, using integration sites that are represented at least ten times in the sequencing data[49].

Sample size. No statistical methods were used to predetermine sample size.

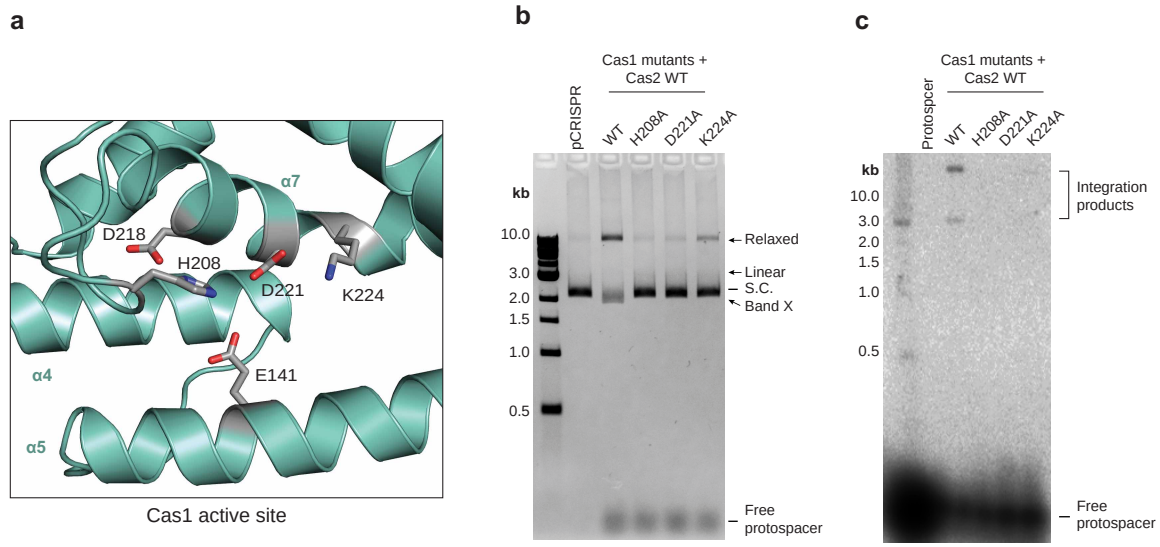
47. Engelman, A. & Craigie, R. Efficient magnesium-dependent human immunodeficiency virus type 1 integrase activity. *J. Virol.* **69,** 5908–5911 (1995).
48. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).
49. Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14,** 1188–1190 (2004).



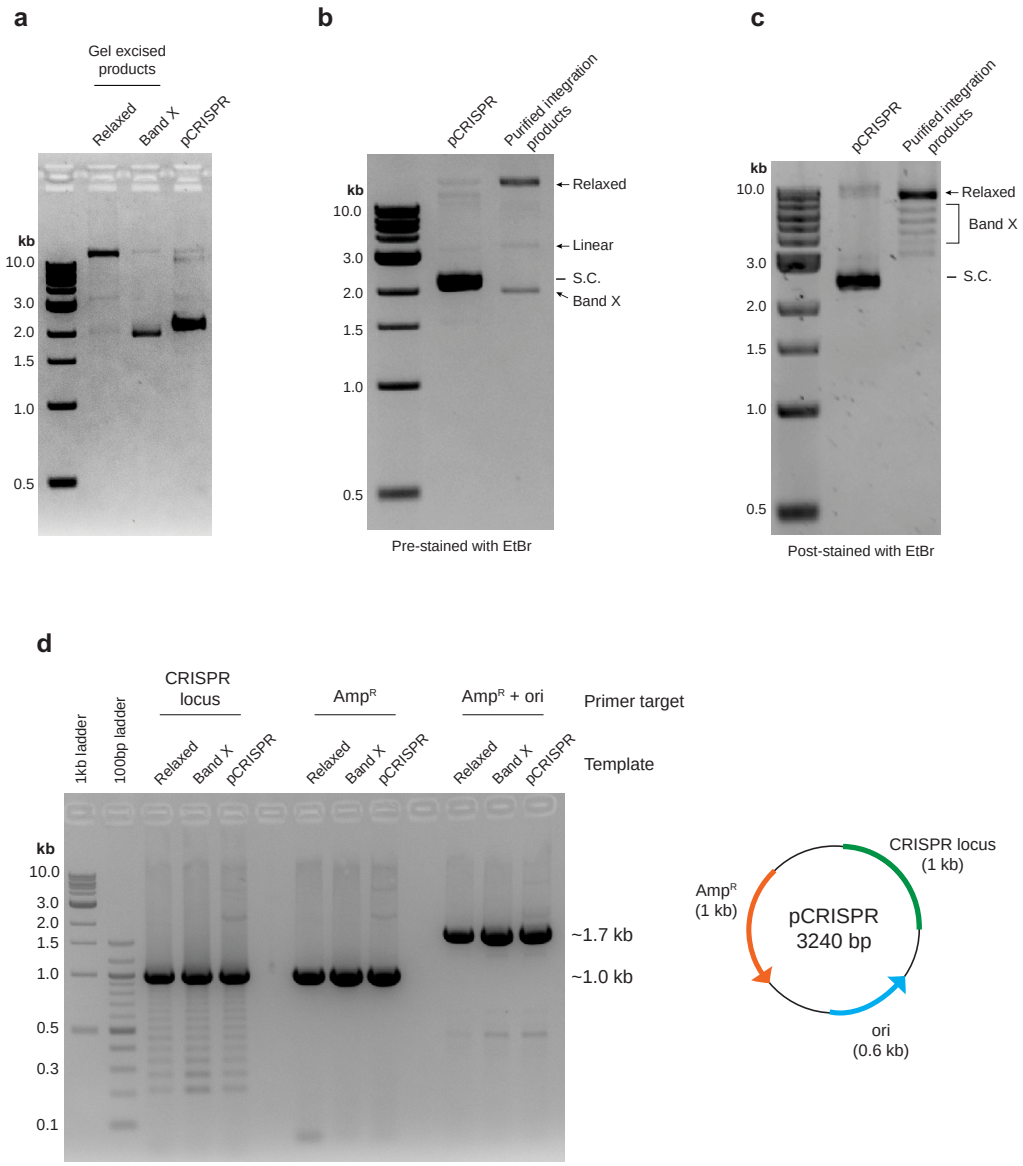
Extended Data Figure 1 | The integration reaction is dependent on the presence of protospacers, low salt and divalent metal ions. **a**, *In vitro* integration assay alongside EcoRI- and Nb.BbvCI nickase-treated pCRISPR. **b**, Salt-dependence assay using Cas1 or Cas2 only and Cas1+Cas2. The titration corresponds to 0, 25, 50, 100 and 200 nM KCl, in addition to the salt carried in from the reaction reagents. **c**, Integration assays in the presence of 10 mM EDTA, Mg^{2+}, Mn^{2+} or no additive. **d**, Integration assays with increasing protospacer concentrations. **e**, A comparison of post-reaction treatments as indicated. The data presented in **a**–**e** are representative of at least three replicates.



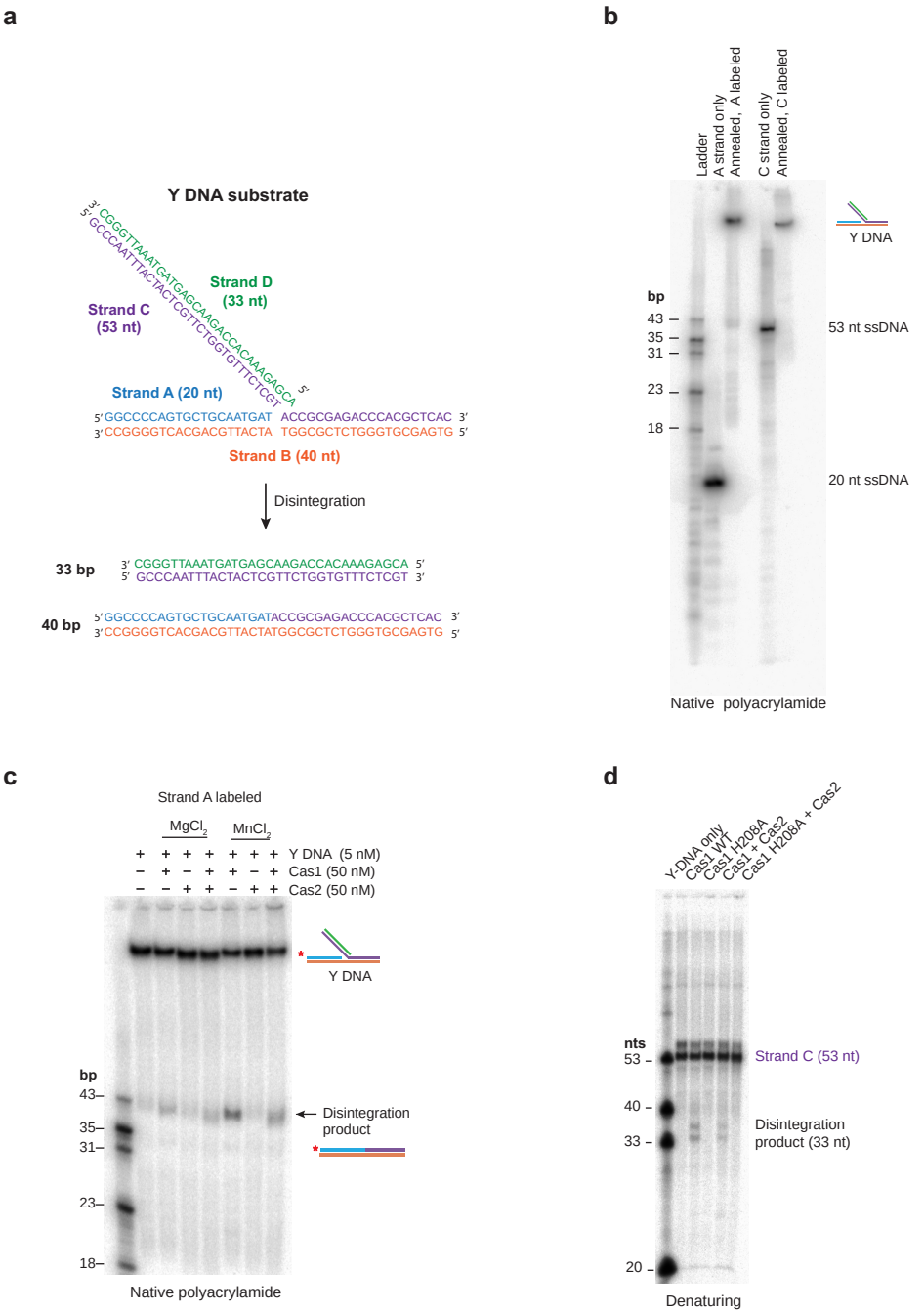
Extended Data Figure 2 | Cas1 requires Cas2 for robust protospacer integration. **a**, Schematic of the integration assays using ^{32}P-labelled protospacers (PDB code 4P6I for Cas1–Cas2). **b**, Integration assays in the presence of increasing protein and 10 mM $MnCl_2$. The titration corresponds to 0, 50, 100 and 200 nM protein. **c**, Same as **b** except in the presence of 10 mM $MgCl_2$. The data presented in **b** and **c** are representative of at least three replicates.



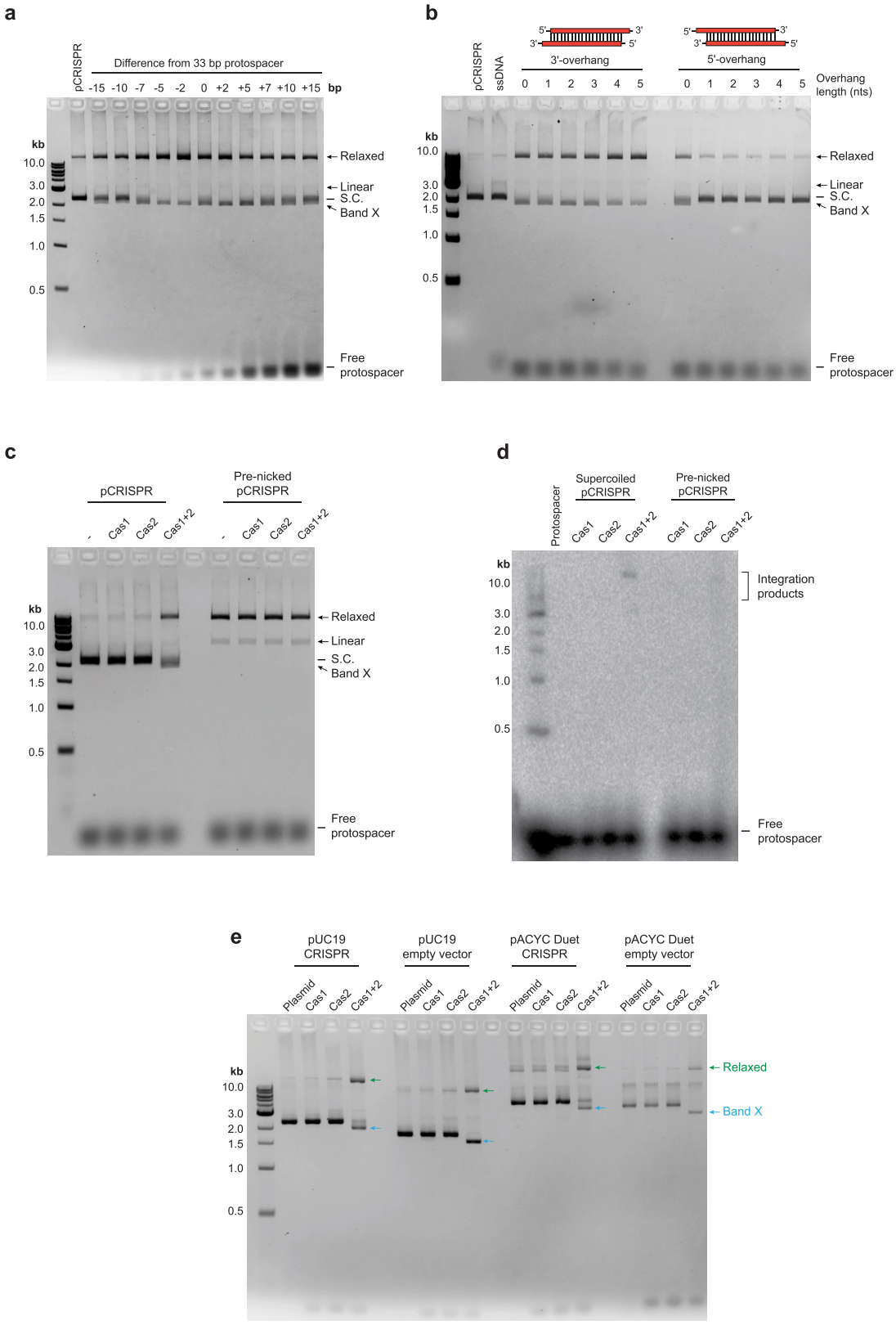
Extended Data Figure 3 | The catalytic activity of Cas1 is required for integration. **a**, Close-up view of the Cas1 active site with the conserved residues shown in stick configurations (PDB 4P6I). **b**, Integration assays of purified Cas1 active site mutants complexed with wild-type Cas2. **c**, The same as **b** except using radiolabelled protospacers. The data presented in **b** and **c** are representative of at least three replicates.



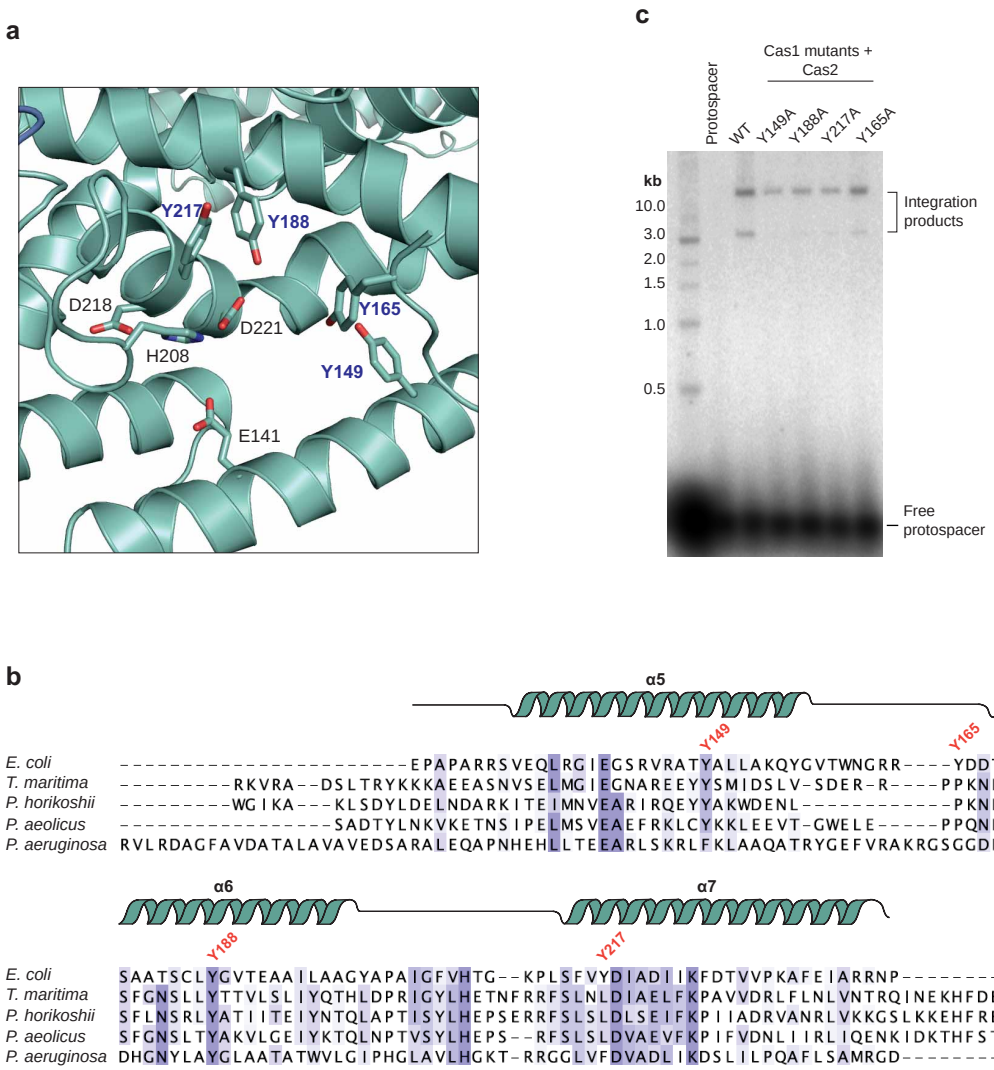
Extended Data Figure 4 | Band X corresponds to topoisomers of pCRISPR. **a**, Agarose gel of purified relaxed and band X integration products. **b**, Analysis of the total reaction products, after phenol chloroform extraction and ethanol precipitation, on a pre-stained agarose gel. **c**, Same as **b** except ethidium bromide staining was performed after electrophoresis. **d**, PCR amplification products of various segments of pCRISPR using the relaxed, band X or pCRISPR template shown in **a**. The laddering effect of minor products using CRISPR locus primers likely reflects the propensity of CRISPR repeats to form DNA hairpins. The data presented in **a**–**d** are representative of at least three replicates.



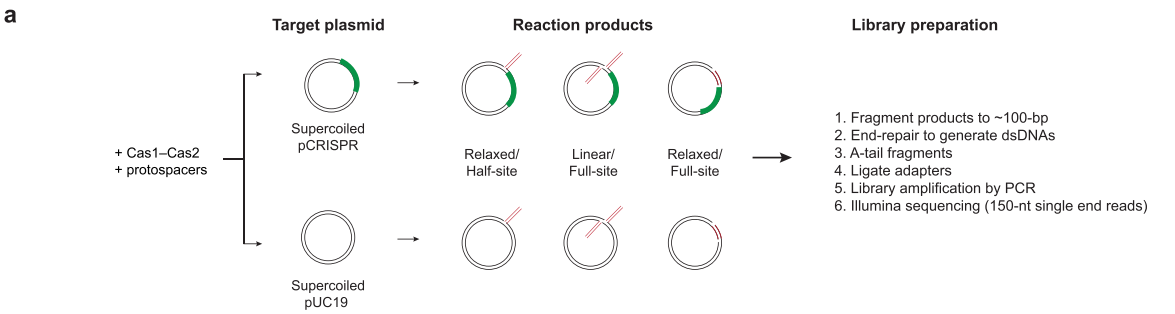
Extended Data Figure 5 | Cas1 catalyses the disintegration of half-site integrated protospacers. **a**, Schematic of the four strands constituting the Y DNA substrate used in the disintegration assays. **b**, Native polyacrylamide gel analysis of the annealing products with either strand A or strand C radiolabelled. **c**, Native polyacrylamide gel analysis of disintegration assay products using Y DNA substrates with strand A labelled. **d**, Denaturing gel analysis of the disintegration assay products with strand A labelled.

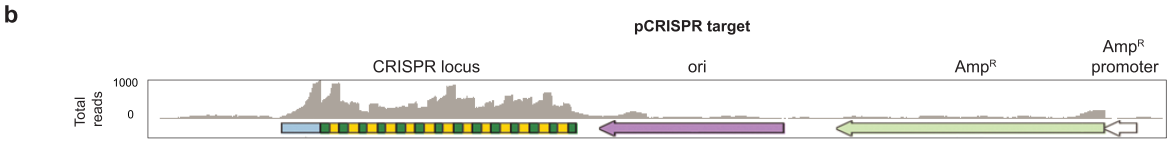


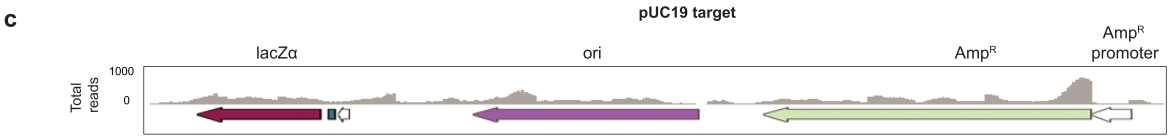
Extended Data Figure 6 | Cas1–Cas2 can integrate various lengths of double-stranded DNA with blunt- or 3′-overhang ends into a supercoiled target plasmid. **a**, Integration assays using the indicated lengths of protospacer DNA. **b**, Integration assays using varying 5′ or 3′ overhang lengths. **c**, **d**, A comparison of integration assays using pCRISPR or Nb.BbvCI-nicked pCRISPR target. **e**, Integration assay using different target plasmids with or without a CRISPR locus. The green arrows correspond to the relaxed product of each target and the cyan arrows correspond to the band X product. The data presented in **a**–**e** are representative of at least three replicates.



Extended Data Figure 7 | Cas1 tyrosine mutants support integration activity *in vitro*. **a**, A close-up of the Cas1 active site with the tyrosine residues labelled in blue. **b**, Structure-based sequence alignment of Cas1 proteins, highlighting the tyrosine residues mutated to alanine in this study. **c**, Radiolabelled protospacer integration assay of Cas1 tyrosine mutants complexed with wild-type Cas2. The gel presented in **c** is representative of at least three replicates.



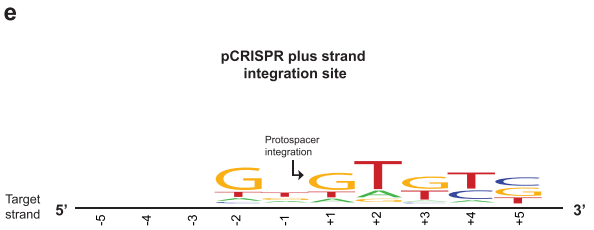


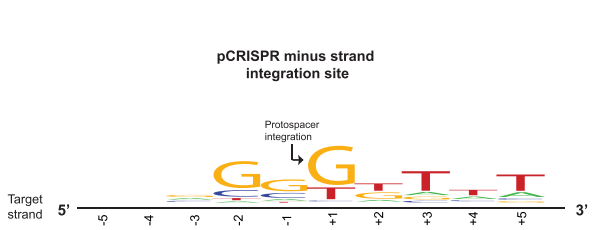


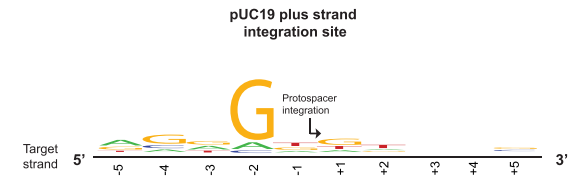
d

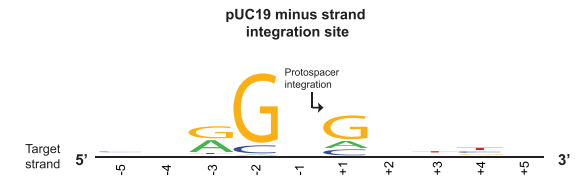
Strand		
	Leader	Repeat-1
+	5' ...TTTAGA	GTGTTCCCCGCGCCAGCGGGGATAAACC
−	3' ...AAATCT	CACAAGGGGCGCGGTCGCCCCTATTTGG

Spacer-1	Repeat-2	Spacer-2
GAGCACAAATATCATCGCTCAAACCACTTACGG	GTGTTCCCCGCGCCAGCGGGGATAAACC	GCCTC... 3'
CTCGTGTTTATAGTAGCGAGTTTGGTGAATGCC	CACAAGGGGCGCGGTCGCCCCTATTTGG	CGGAG...5'

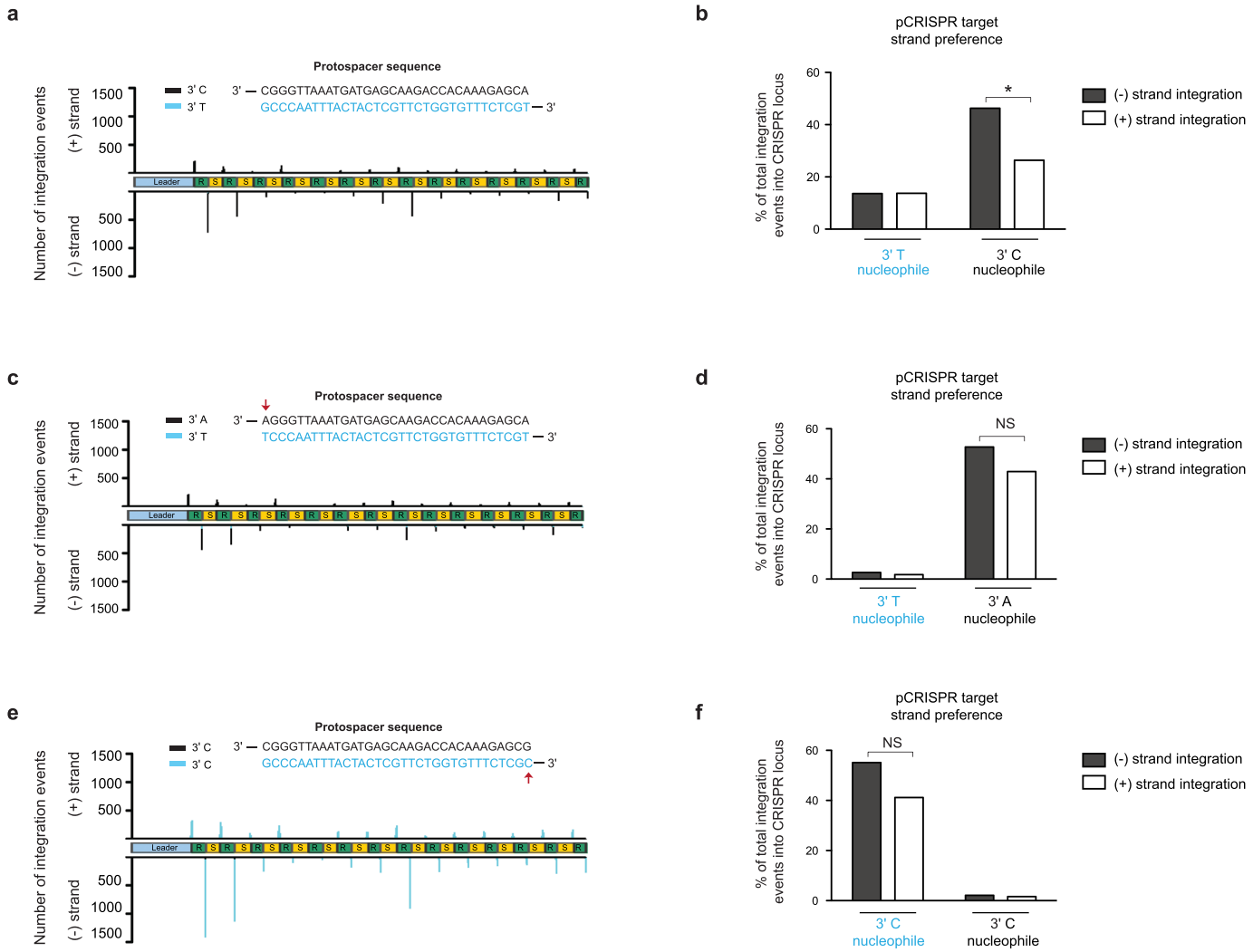




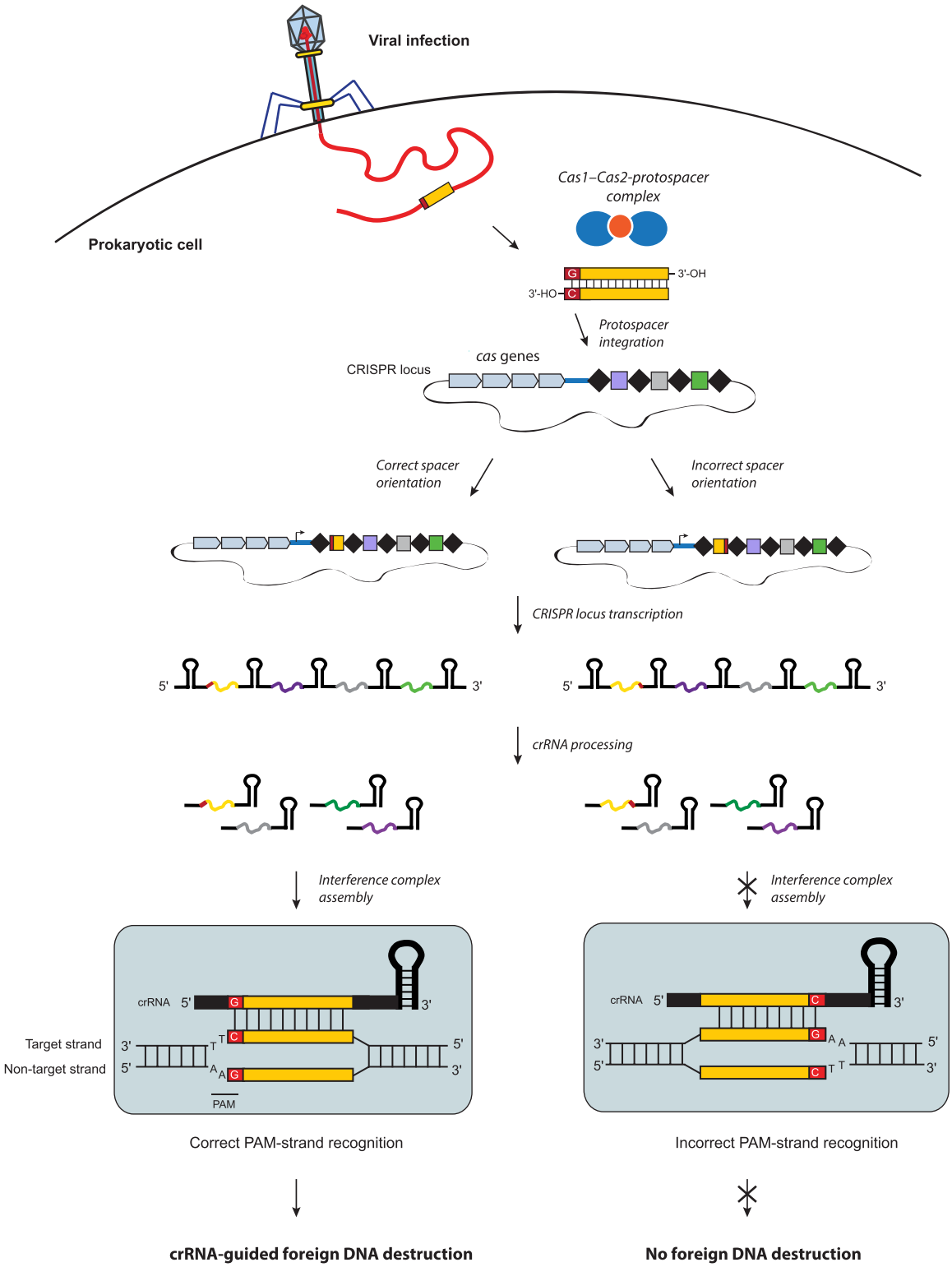




Extended Data Figure 8 | High-throughput sequencing of integration products reveals sequence-specific integration. **a**, Schematic of the workflow for high-throughput sequencing analysis of the integration sites. **b**, Raw map of the total reads along pCRISPR before collapsing into single peaks of protospacer–pCRISPR junctions depicted in Fig. 4. **c**, Same as **b**, except for the pUC19 target. **d**, Sequence of the leader-end of the CRISPR locus in *E. coli*. **e**, **f**, WebLogo analysis from the −5 to +5 positions surrounding the protospacer integration sites on the plus (**e**) and minus (**f**) of pCRISPR. The arrow points to the nucleotide that is covalently joined to the protospacer. **g**, **h**, Same as **e**, **f**, except for the pUC19 target.



Extended Data Figure 9 | Cas1–Cas2 correctly orients the protospacer DNA during integration. **a**–**f**, Mapped integration sites along the CRISPR locus of pCRISPR when using protospacer DNA with nucleotide ends 'wild-type' 3′ C and 3′ T (**a**), 3′ A and 3′ T (**c**), and 3′ C and 3′ C (**e**). The red arrow in **c** and **e** points to the nucleotide change in the protospacer DNA compared to the 'wild-type' sequence in **a**. The protospacer DNA 3′ nucleotide and the CRISPR locus strand biases in **a**, **c**, **e** are plotted in **b**, **d** and **f**, respectively, as percentages of integration events within the CRISPR locus. The black and clear bars represent the (−) and (+) strands of the CRISPR locus, respectively. NS corresponds to not significant and $*P < 0.0001$ by chi-square test. The *n* values for **b**, **d** and **f** are 5,623, 5,685 and 12,453 reads along the CRISPR locus, respectively.



Extended Data Figure 10 | Model of the CRISPR–Cas adaptive immunity pathway in *E. coli*. Mature double-stranded protospacers bearing a 3′ C-OH are site-specifically integrated into the leader-end of the CRISPR locus. Correct protospacer integration (left) results in the 5′G/3′C as the first nucleotide of the spacer, proximal to the leader. After transcription of the CRISPR locus and subsequent crRNA processing, foreign DNA destruction is initiated by strand-specific recognition of the 3′-TTC-5′ PAM sequence in the target strand by the crRNA-guided Cascade complex. Incorrect protospacer integration (right) cannot initiate foreign DNA destruction due to the inability for the crRNA to recognize the strand with the 3′-TTC-5′ PAM. Thus, foreign DNA interference during CRISPR–Cas adaptive immunity relies on the Cas1–Cas2 complex for correctly orienting the protospacer during integration.

ARTICLE

doi:10.1038/nature14245

Cas9 specifies functional viral targets during CRISPR–Cas adaptation

Robert Heler[1]*, Poulami Samai[1]*, Joshua W. Modell[1], Catherine Weiner[1], Gregory W. Goldberg[1], David Bikard[1,2] & Luciano A. Marraffini[1]

Clustered regularly interspaced short palindromic repeat (CRISPR) loci and their associated (Cas) proteins provide adaptive immunity against viral infection in prokaryotes. Upon infection, short phage sequences known as spacers integrate between CRISPR repeats and are transcribed into small RNA molecules that guide the Cas9 nuclease to the viral targets (protospacers). *Streptococcus pyogenes* Cas9 cleavage of the viral genome requires the presence of a 5′-NGG-3′ protospacer adjacent motif (PAM) sequence immediately downstream of the viral target. It is not known whether and how viral sequences flanked by the correct PAM are chosen as new spacers. Here we show that Cas9 selects functional spacers by recognizing their PAM during spacer acquisition. The replacement of *cas9* with alleles that lack the PAM recognition motif or recognize an NGGNG PAM eliminated or changed PAM specificity during spacer acquisition, respectively. Cas9 associates with other proteins of the acquisition machinery (Cas1, Cas2 and Csn2), presumably to provide PAM-specificity to this process. These results establish a new function for Cas9 in the genesis of prokaryotic immunological memory.

CRISPR loci and Cas proteins provide adaptive immunity to bacteria and archaea against their viruses[1]. To adapt to highly dynamic viral populations, CRISPR–Cas loci evolve rapidly, acquiring short phage sequences, known as spacers, that integrate between CRISPR repeats and constitute a memory record of infection[2]. Spacers are transcribed into small CRISPR RNAs (crRNAs) that identify viral targets (defined as protospacers) by direct Watson–Crick pairing with invasive DNA[3]. Based on their *cas* gene content, CRISPR–Cas systems can be classified into three distinct types: I, II and III (ref. 4). Each CRISPR–Cas type possesses different mechanisms of crRNA biogenesis, target destruction and prevention of autoimmunity. In the type II CRISPR–Cas system present in *Streptococcus pyogenes* the Cas9 nuclease inactivates infective phages using crRNAs as guides to introduce double-strand DNA breaks into the viral genome[5]. Cas9 cleavage requires the presence of a protospacer adjacent motif (PAM) sequence immediately downstream of the protospacer[6,7]. This requirement avoids the cleavage of the spacer sequence within the CRISPR array, that is, autoimmunity, as the adjacent repeat lacks a PAM sequence. The importance of the PAM sequence for target recognition and cleavage[6–9] suggests the presence of a mechanism to ensure that newly acquired spacer sequences match protospacers flanked by a proper PAM sequence. For the type I-E CRISPR–Cas system of *Escherichia coli*, overexpression of *cas1* and *cas2* is sufficient for the acquisition of new spacers in the absence of phage infection. Reports indicate that spacers acquired in this fashion match preferentially (25–70%, depending on the study) to protospacers with the correct PAM (AWG, W = A/T)[10–13], suggesting that Cas1 and Cas2 are sufficient for spacer acquisition and have some intrinsic ability to recognize protospacers with the right PAM. In the type II system of *S. pyogenes*, the PAM sequence is NGG (and also NAG at a much lower frequency)[3,6,14], where N is any nucleotide; this motif is recognized and bound by a domain within the Cas9 nuclease during target cleavage[7,15]. How spacers are acquired in this system, particularly how spacers with correct PAM sequences are selected during this process, is not known.

Cas9 is required for spacer acquisition

To investigate the mechanisms of recognition of PAM-adjacent protospacers during spacer acquisition, we cloned the type II-A CRISPR–Cas locus of *S. pyogenes* (Fig. 1a) into the staphylococcal vector pC194 (ref. 16) and introduced the resulting plasmid (pWJ40 (ref 17)) into *Staphylococcus aureus* RN4220 (ref. 18), a strain lacking CRISPR–Cas loci. We chose this experimental system because it facilitates the genetic manipulation of the *S. pyogenes* CRISPR–Cas system. We first tested the ability of the cells to mount adaptive CRISPR immunity by infecting

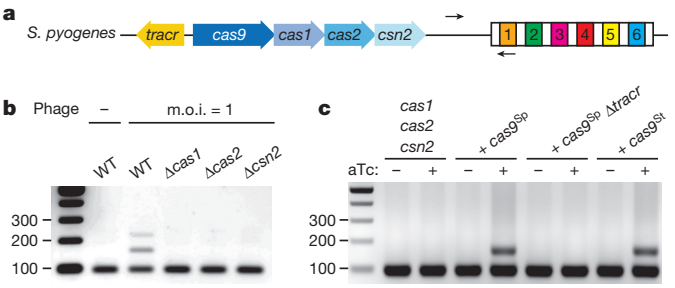


Figure 1 | Cas9 is required for spacer acquisition. **a**, Organization of the *S. pyogenes* type II CRISPR–Cas locus. Arrows indicate the annealing position of the primers used to check for the expansion of the CRISPR array. **b**, PCR-based analysis of liquid cultures to check for the acquisition of new spacer sequences in the presence or the absence of phage ϕNM4γ4 infection. Wild type (WT) as well as different *cas* mutants were analysed. Image is representative of three technical replicates. m.o.i., multiplicity of infection. **c**, Cultures overexpressing Cas1, Cas2 and Csn2 under the control of a tetracycline-inducible promoter were analysed using PCR for spacer acquisition in the absence of phage infection. The strain was complemented with plasmids carrying either *Streptococcus thermophilus* (St) or *S. pyogenes* (Sp) Cas9 (see Extended Data Fig. 3), in the last case with or without the tracrRNA gene (Δ*tracr*). Image is representative of three technical replicates. aTc, anhydrotetracycline.

[1]Laboratory of Bacteriology, The Rockefeller University, 1230 York Avenue, New York, New York 10065, USA. [2]Synthetic Biology Group, Institut Pasteur, 28 Rue du Dr. Roux, 75015 Paris, France.
*These authors contributed equally to this work.

them with the staphylococcal phage ϕNM4γ4, a lytic variant of ϕNM4 (ref. 19) (see Methods for a description of ϕNM4γ4 isolation). Plate-based assays performed by mixing bacteria and phage in top agar allowed the selection of phage-resistant colonies that were checked by PCR to look for the expansion of the CRISPR array (Extended Data Fig. 1a). On average 50% of the colonies acquired one or more spacers (8/13, 5/11 and 7/16 in three independent experiments), whereas the rest of the resistant colonies survived phage infection by a non-CRISPR mechanism, most likely including phage receptor mutations (Extended Data Fig. 2a). To maximize the capture of new spacer sequences, we performed the same assay in liquid and recovered surviving bacteria at the end of the phage challenge. These were analysed by PCR of the CRISPR array and the amplification products of expanded loci were subjected to Illumina MiSeq sequencing to determine the extent of spacer acquisition. Analysis of 2.96-million reads detected protospacers adjacent to 2,083 out of 2,687 NGG sequences present in the viral genome, although with variation in the frequency of acquisition of each sequence (Extended Data Fig. 1b). The data revealed a prominent selection of spacers matching protospacers with downstream NGG PAM sequences (99.97%, Extended Data Fig. 1c). The acquisition of new spacers by cells in liquid culture proved to be simple and highly efficient, providing the possibility to look at millions of new spacers in a single step. It was therefore used in the rest of our studies.

To determine the genetic requirements for spacer acquisition, we made individual deletions of *cas1*, *cas2* or *csn2* and challenged the mutant strains with phage ϕNM4γ4. Spacer acquisition was decreased to levels below our limit of detection in each of these mutants (Fig. 1b), corroborating previous experiments[12,20]. Therefore although Cas1, Cas2 and Csn2 are dispensable for anti-phage immunity in the presence of a pre-existing spacer (Extended Data Fig. 2b, c), they are required for spacer acquisition. To determine whether these genes are also sufficient for this process, we overexpressed *cas1*, *cas2* and *csn2* in the absence of *cas9* using a tetracycline-inducible promoter in plasmid pRH223 and looked for the integration of new spacers in the absence of phage infection using a highly sensitive PCR assay (Extended Data Fig. 3). We were unable to detect new spacers even in the presence of the inducer (Fig. 1c). However, the addition of a second plasmid expressing tracrRNA (see below) and Cas9 from their native promoters (Extended Data Fig. 3) enabled spacer acquisition only in the presence of the inducer, with all the new spacers matching chromosomal or plasmid sequences (Fig. 1c and Extended Data Table 1). Although it is most likely that the acquisition of such spacers causes cell death or plasmid curing, respectively, the acquisition event can still be detected in liquid culture using our highly sensitive PCR assay (Extended Data Fig. 3b, c). The tracRNA (Fig. 1a) is a small RNA bound by Cas9 that is required for crRNA processing[3] and Cas9 nuclease activity[6]. We wondered if Cas9 involvement in spacer acquisition also required the presence of the tracrRNA. Deletion of the tracrRNA prevented spacer acquisition in the absence of phage infection (Fig. 1c), suggesting that apo-Cas9 is not sufficient to promote spacer acquisition and that association with its cofactor is also required. Altogether these data indicate that Cas1, Cas2 and Csn2 are necessary but not sufficient for the incorporation of new spacers and that a tracrRNA–Cas9 complex is also required. This is in contrast to the type I-E CRISPR–Cas system of *E. coli*, in which overexpression of Cas1 and Cas2 alone is sufficient for spacer acquisition[10–13]. It is important to note that the CRISPR array used in this assay consists of a single repeat, without pre-existing spacers (Extended Data Fig. 3). Therefore the Cas9 requirement is not a consequence of the phenomenon known as 'primed' spacer acquisition. This refers to an increase in the frequency of spacer acquisition observed in type I CRISPR–Cas systems that relies on the presence of a pre-existing spacer with a partial match to the phage genome as well as the full targeting complex (Cascade)[12,21,22].

Cas9 specifies the PAM of newly acquired spacers

Given this newfound requirement in the CRISPR adaptation process and the well-established PAM recognition function of Cas9 during the surveillance and destruction of viral target sequences, we hypothesized that this nuclease could participate in the selection of PAM sequences during spacer acquisition. To test this we exchanged the *cas9* genes of *S. pyogenes* (Sp) and *S. thermophilus* (St) CRISPR–Cas systems to create two chimaeric CRISPR loci: *tracrRNA*^{Sp}-*cas9*St-*cas1*^{Sp}-*cas2*^{Sp}-*csn2*^{Sp} and *tracrRNA*St-*cas9*^{Sp}-*cas1*St-*cas2*St-*csn2*St (Fig. 2a). We chose the type II-A CRISPR–Cas system of *S. thermophilus* (also known as CRISPR3 (ref. 23)) because it is an orthologue of the *S. pyogenes* system[24]. While the PAM sequence for the Sp CRISPR–Cas system is NGG, the PAM sequence for the St system is NGGNG[23] (Fig. 2b and Extended Data Table 1). We infected each naive strain with phage ϕNM4γ4, sequenced the newly acquired spacers, and obtained the PAM of the matching protospacers using WebLogo[25]. We found that each chimaeric system acquired spacers with PAMs that correlated with the *cas9*, but not the *tracrRNA*, *cas1*, *cas2* or *csn2*, allele present (Fig. 2b and Extended Data Table 1). To rule out the possibility that non-functional spacers are negatively selected during phage infection, that is, they are acquired randomly but only those cells containing spacers with a correct PAM for Cas9 cleavage provide immunity and allow cell survival, we sequenced the PAMs of spacers acquired in the absence of phage infection (Figs 1c and 2c). Either Cas9^{Sp} or Cas9St were produced in cells overexpressing Cas1^{Sp}, Cas2^{Sp} and Csn2^{Sp}. In this experiment, as explained earlier, spacers matching chromosomal or plasmid sequences were acquired. The PCR products containing new spacers were cloned into a commercial vector from which they were sequenced (Extended Data Table 1). Expression of Cas9^{Sp} led to the incorporation of spacers matching protospacers with an NGG PAM sequence, whereas the expression of Cas9St in the same cells shifted the composition of the PAM to NGGNG (Fig. 2d). These results demonstrate that Cas9 determines PAM sequences during CRISPR adaptation to ensure the acquisition of functional spacers.

Cas9 associates with Cas1, Cas2 and Csn2

In type I CRISPR–Cas systems, Cas1 and Cas2 form a complex[13] and the dsDNA nuclease activity of Cas1 has been implicated in the initial cleavage of the invading viral DNA to generate a new spacer[26]. The genetic analyses presented above suggest that in the type II *S. pyogenes* CRISPR–Cas system, the PAM-binding function of Cas9 observed *in vitro*[7] could specify a PAM-adjacent site of cleavage for Cas1, or other members of the spacer acquisition machinery. This would guarantee that newly acquired spacers have the correct PAM needed for Cas9 activity later in this immune pathway. This hypothesis predicts an interaction between Cas9 and Cas1, Cas2 and/or Csn2. To test this we expressed the type II Cas operon in *E. coli*, using a histidyl tagged version of Cas9, and looked for other proteins that co-purify. We observed an abundant co-purifying protein with an apparent molecular weight close to 33 kDa, the expected size of Cas1 (Extended Data Fig. 4a). Mass spectrometry confirmed the identity of both of these proteins, as well as the presence of Cas2 and Csn2 co-purifying with Cas9 (Extended Data Table 2). This result suggested the formation of a Cas9–Cas1–Cas2–Csn2 complex and therefore we explored other purification strategies to unequivocally determine its existence. We were able to isolate a Cas9–Cas1–Cas2–Csn2 complex when the histidyl tag was added to Csn2 (Fig. 3a, b). The identity of the purified proteins was confirmed by mass spectrometry (Extended Data Table 3). This demonstrates a biochemical link between the Cas9 nuclease and the other Cas proteins that function exclusively to acquire new spacers, supporting the role of Cas9 as a PAM specificity factor in the adaptation phase of CRISPR immunity.

Cas9 PAM-binding motif is needed for spacer selection

Within this complex the PAM-binding domain of Cas9 would specify a functional spacer (one adjacent to a correct PAM) and the nuclease activity of Cas1 and/or Cas9 would cleave the invading DNA to extract the spacer sequence. To test this model, we performed adaptation studies in the absence of phage selection as described in Extended Data Fig. 3 but using different combinations of wild-type Cas1, Cas1(E220A) (catalytically dead or dCas1 (ref. 26)), wild-type Cas9, Cas9(D10A,H840A)

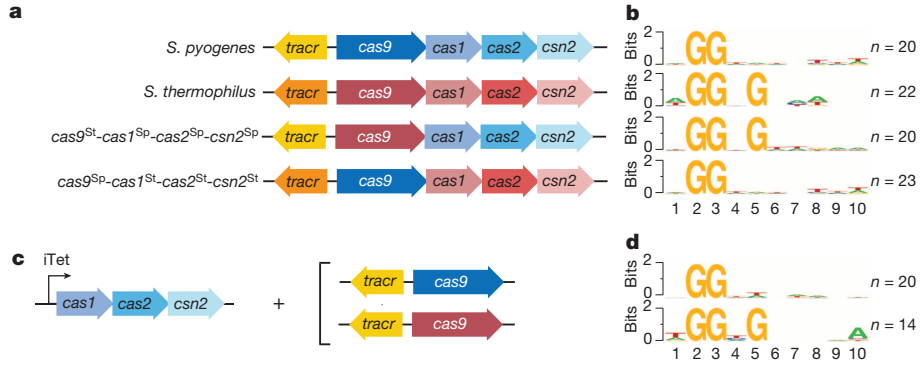


Figure 2 | Cas9 determines the PAM sequence of acquired spacers. **a**, **c**, Genetic composition of the CRISPR–Cas loci tested for spacer during phage infection (**a**), or in the absence of infection (**c**), with the experimental set up shown in Extended Data Fig. 3. **b**, **d**, Sequence logos obtained after the alignment of the 3′ flanking sequences of the protospacers matched by the newly acquired spacers in panels **a** and **c**, respectively. Numbers indicate the positions of the flanking nucleotides downstream from the spacer. Number of sequences used in each alignment indicated as *n*.

(catalytically dead or dCas9 (ref. 6)) and Cas9(R1333Q,R1335Q) (abbreviated here as Cas9^{PAM}, containing mutations in the PAM-binding motif that substantially reduces binding to target DNA sequences with NGG PAMs *in vitro*[15]). We observed that the nuclease activity of Cas1 is necessary for spacer acquisition (Fig. 3c). In contrast, the nuclease activity and PAM-binding function of Cas9 are dispensable for this process. Next we determined the PAM of the acquired spacers in the presence of mutated Cas9 (Fig. 3d). We found that whereas spacers acquired in the presence of dCas9 displayed correct PAMs, those acquired in the presence of Cas9^{PAM} matched DNA regions without a conserved flanking sequence, that is, without a PAM sequence. Cells containing the catalytically dead Cas9(D10A,H847A) from *S. thermophilus* acquired spacers with NGGNG PAMs (Extended Data Fig. 5). These results indicate that Cas1 and Cas9 are part of a complex dedicated to spacer acquisition which requires Cas1 nuclease activity and Cas9 PAM-binding properties for the selection of new spacer sequences.

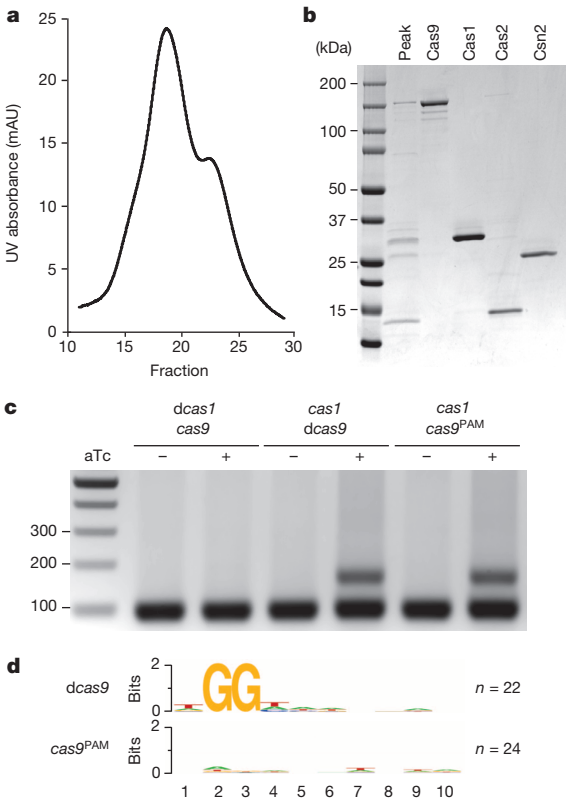


Figure 3 | *S. pyogenes* Cas9 PAM recognition domain is required for the acquisition of spacers with an NGG PAM sequence. **a**, Separation of the Cas9–Cas1–Cas2–Csn2 complex by ion exchange chromatography. **b**, SDS–PAGE of fraction 19 (peak) from the complex elution shown in panel **a**, representative of five technical replicates. The four proteins of the complex were individually purified and run alongside the purified fraction to identify each protein in the complex. **c**, Spacer acquisition was tested as in Fig. 1c in the presence or absence of different Cas1 or Cas9 activities. Image is representative of eight technical replicates. dCas1, nuclease-dead Cas1 (E220A mutation); dCas9, nuclease-dead Cas9 (D10A, H840A mutations); Cas9^{PAM} lacks the PAM recognition function (R1333Q, R1335Q mutations). **d**, Sequence logos obtained after the alignment of the 3′ flanking sequences of the protospacers matched by the newly acquired spacers in panel **c**. Numbers indicate the positions of the flanking nucleotides downstream from the spacer. Number of sequences used in each alignment indicated as *n*.

Discussion

The selection of new spacers with a correct PAM is fundamental for the survival of the infected host during CRISPR–Cas immunity. In the simplest scenario there is no active selection of PAM-flanked protospacers; any spacer sequence can be acquired but only those with the correct PAM allow Cas9 cleavage of the invader and survival. Bacteria that acquire spacers with ineffective flanking sequences are killed by the virus and as a consequence PAM-flanking spacers are enriched in the population. Here we show that even in the absence of phage selection, the type II CRISPR–Cas system acquires new spacers with correct PAMs, a result that rules out the possibility of random spacer selection with subsequent selection for functional spacers. How are PAM-flanked protospacers selected during type II CRISPR–Cas immunity? One possibility is that the proteins exclusively dedicated to spacer acquisition perform the PAM-selection function. The inability of cells overexpressing only *cas1*, *cas2* and *csn2* to expand the CRISPR array strongly suggest that none of the proteins encoded by these genes can recognize and select correct PAMs. Another possibility is that the known PAM-recognition function of Cas9 (refs 15, 27), essential for destroying the invading virus, could also be used during spacer acquisition to recognize PAM-flanking viral sequences. Experiments showing that the *cas9* allele, but not the *cas1* or *cas2* or *csn2* alleles, determines the PAM sequence of the newly acquired spacers, demonstrated that this scenario is probably correct. Regarding the molecular mechanism by which Cas9 participates in CRISPR adaptation, our experiments show that Cas9 forms a stable complex with Cas1, Cas2 and Csn2 that presumably participates in the selection of new spacers. The nuclease activity of Cas1, but not of Cas9, is required for spacer acquisition. The tracrRNA is also required, suggesting that the apo-Cas9 structure[27], very different from holo-Cas9 (ref. 15), does not have the correct conformation to participate in spacer acquisition. The key residues involved in Cas9 PAM recognition are not required for spacer acquisition, but they are necessary for the incorporation of new spacers with the correct PAM sequence. This suggests that the reported non-specific DNA binding

property of Cas9 (refs 6, 7) is sufficient for spacer acquisition, but not for the selection of functional spacers. There are currently two models for the incorporation of new spacers into the CRISPR array, one where the future spacer sequence is cut from the invading viral DNA, the 'cut and paste' model, and another where this sequence is copied from the viral genome, the 'copy and paste' model[28]. In the context of the first model, our data suggests that, at a low frequency that may reflect the dynamics of spacer acquisition, Cas1 cleaves the invading genome to extract a new spacer sequence. However, on its own, Cas1 nuclease activity is non-specific[26]. Therefore we propose that through the formation of the Cas9–Cas1–Cas2–Csn2 complex, Cas9 binding to PAM-adjacent sequences provides specificity to Cas1 endonuclease activity. In the copy and paste model, Cas1 nuclease activity is most likely necessary for downstream events, such as the cleavage of the repeat sequence that precedes spacer insertion, and Cas9 is required to 'mark' sequences adjacent to GG motifs to be copied into the CRISPR array. In any case, following as yet unknown processing and integration events, the selected DNA becomes a new functional spacer, that is, its matching protospacer will have the correct PAM to license Cas9 cleavage (Extended Data Fig. 6). The molecular steps that take place after protospacer selection to incorporate it as a new spacer in the CRISPR array are still unknown. All genes of the type II-A CRISPR–Cas locus (tracrRNA, *cas9*, *cas1*, *cas2* and *csn2*) are required for spacer acquisition, therefore most likely all the members of the Cas9–Cas1–Cas2–Csn2 complex participate in the process. Future work will address this and other aspects of the mechanisms of spacer integration in different CRISPR–Cas systems. The present work reveals a new function for Cas9 in CRISPR immunity. This nuclease is fundamental for both the execution of immunity, participating in the surveillance and destruction of infectious target viruses, and the generation of immunological memory, selecting the viral sequences that allow adaptation and resistance to viral predators.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 15 October 2014; accepted 20 January 2015.
Published online 18 February 2015.

1. Barrangou, R. & Marraffini, L. A. CRISPR–Cas systems: prokaryotes upgrade to adaptive immunity. *Mol. Cell* **54,** 234–244 (2014).
2. Barrangou, R. *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315,** 1709–1712 (2007).
3. Deltcheva, E. *et al.* CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471,** 602–607 (2011).
4. Makarova, K. S. *et al.* Evolution and classification of the CRISPR–Cas systems. *Nature Rev. Microbiol.* **9,** 467–477 (2011).
5. Garneau, J. E. *et al.* The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468,** 67–71 (2010).
6. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337,** 816–821 (2012).
7. Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C. & Doudna, J. A. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507,** 62–67 (2014).
8. Gasiunas, G., Barrangou, R., Horvath, P. & Siksnys, V. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc. Natl Acad. Sci. USA* **109,** E2579–E2586 (2012).
9. Szczelkun, M. D. *et al.* Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes. *Proc. Natl Acad. Sci. USA* **111,** 9798–9803 (2014).
10. Díez-Villaseñor, C., Guzman, N. M., Almendros, C., Garcia-Martinez, J. & Mojica, F. J. CRISPR-spacer integration reporter plasmids reveal distinct genuine acquisition specificities among CRISPR–Cas I-E variants of *Escherichia coli*. *RNA Biol.* **10,** 792–802 (2013).
11. Yosef, I., Goren, M. G. & Qimron, U. Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. *Nucleic Acids Res.* **40,** 5569–5576 (2012).
12. Datsenko, K. A. *et al.* Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. *Nat. Commun.* **3,** 945 (2012).
13. Nuñez, J. K. *et al.* Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR–Cas adaptive immunity. *Nature Struct. Mol. Biol.* **21,** 528–534 (2014).
14. Jiang, W., Bikard, D., Cox, D., Zhang, F. & Marraffini, L. A. RNA-guided editing of bacterial genomes using CRISPR–Cas systems. *Nature Biotechnol.* **31,** 233–239 (2013).
15. Anders, C., Niewoehner, O., Duerst, A. & Jinek, M. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature* **513,** 569–573 (2014).
16. Horinouchi, S. & Weisblum, B. Nucleotide sequence and functional map of pC194, a plasmid that specifies inducible chloramphenicol resistance. *J. Bacteriol.* **150,** 815–825 (1982).
17. Goldberg, G. W., Jiang, W., Bikard, D. & Marraffini, L. A. Conditional tolerance of temperate phages via transcription-dependent CRISPR–Cas targeting. *Nature* **514,** 633–637 (2014).
18. Kreiswirth, B. N. *et al.* The toxic shock syndrome exotoxin structural gene is not detectably transmitted by a prophage. *Nature* **305,** 709–712 (1983).
19. Bae, T., Baba, T., Hiramatsu, K. & Schneewind, O. Prophages of *Staphylococcus aureus* Newman and their contribution to virulence. *Mol. Microbiol.* **62,** 1035–1047 (2006).
20. Sapranauskas, R. *et al.* The *Streptococcus thermophilus* CRISPR/Cas system provides immunity in *Escherichia coli*. *Nucleic Acids Res.* **39,** 9275–9282 (2011).
21. Li, M., Wang, R., Zhao, D. & Xiang, H. Adaptation of the *Haloarcula hispanica* CRISPR–Cas system to a purified virus strictly requires a priming process. *Nucleic Acids Res.* **42,** 2483–2492 (2014).
22. Richter, C. *et al.* Priming in the Type I-F CRISPR–Cas system triggers strand-independent spacer acquisition, bi-directionally from the primed protospacer. *Nucleic Acids Res.* **42,** 8516–8526 (2014).
23. Horvath, P. *et al.* Diversity, activity, and evolution of CRISPR loci in *Streptococcus thermophilus*. *J. Bacteriol.* **190,** 1401–1412 (2008).
24. Fonfara, I. *et al.* Phylogeny of Cas9 determines functional exchangeability of dual-RNA and Cas9 among orthologous type II CRISPR–Cas systems. *Nucleic Acids Res.* **42,** 2577–2590 (2014).
25. Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14,** 1188–1190 (2004).
26. Wiedenheft, B. *et al.* Structural basis for DNase activity of a conserved protein implicated in CRISPR-mediated genome defense. *Structure* **17,** 904–912 (2009).
27. Jinek, M. *et al.* Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science* **343,** 1247997 (2014).
28. Arslan, Z., Hermanns, V., Wurm, R., Wagner, R. & Pul, U. Detection and characterization of spacer integration intermediates in type I-E CRISPR–Cas system. *Nucleic Acids Res.* **42,** 7884–7893 (2014).

Acknowledgements We thank members of the laboratory for critical discussion of the experiments and their results, A. Zaytsev for help with the deep sequencing data analysis and A. Sherlock for help with plasmid construction. R.H. is the recipient of a Howard Hughes International Student Research Fellowship. P.S. is supported by a Helmsley Postdoctoral Fellowship for Basic and Translational Research on Disorders of the Digestive System at The Rockefeller University. J.W.M. is a Fellow of The Jane Coffin Childs Memorial Fund for Medical Research. D.B. is supported by a Harvey L. Karp Discovery Award and the Bettencourt Schuller Foundation. L.A.M. is supported by the Rita Allen Scholars Program, an Irma T. Hirschl Award, a Sinsheimer Foundation Award and a NIH Director's New Innovator Award (1DP2AI104556-01).

Author Contributions R.H., P.S., D.B. and L.A.M. conceived the study and designed experiments. R.H. and P.S. executed the experimental work with help from C.W. J.W.M. set up the experimental system to detect spacer acquisition in the absence of phage infection. G.W.G. isolated and characterized phage ϕNM4γ4 and constructed the pGG32 plasmid. D.B. analysed MiSeq data. L.A.M. wrote the paper with the help of the rest of the authors.

Author Information The full sequence of ϕNM4γ4 has been deposited in GenBank under accession number KP209285. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.B. (david.bikard@pasteur.fr) or L.A.M. (marraffini@rockefeller.edu).

METHODS

Bacterial strains and growth conditions. Cultivation of *S. aureus* RN4220 (ref. 18), was carried out in brain-heart infusion (BHI) or heart infusion (HI) media (BD) at 37 °C. Whenever applicable, media were supplemented with chloramphenicol at 10 μg ml^{-1} or erythromycin at 5 μg ml^{-1} to ensure pC194-derived (ref. 16) and pE194-derived[29] plasmid maintenance, respectively.

On-plate spacer acquisition assay. To detect individual adapted colonies on a plate, cells from overnight cultures were mixed with phage at a m.o.i. value of 1 in top agar containing appropriate antibiotic and 5 mM $CaCl_2$. The mixture was poured on BHI plates with antibiotic and incubated at 37 °C overnight. Subsequently, colonies that survived phage infection were restreaked on fresh BHI plates in order to remove contaminating virus and dead cells. Plates were incubated at 37 °C overnight. To check for spacer acquisition, individual colonies were resuspended in lysis buffer (250 mM KCl, 5 mM $MgCl_2$ 50 mM Tris-HCl at pH 9.0, 0.5% Triton X-100), treated with 50 ng μl^{-1} lysostaphin and incubated at 37 °C for 5 min, then 98 °C for 5 min. Following centrifugation (16,000*g*), a sample of the supernatant was used as template for TopTaq PCR amplification with primers L400 and H050. The PCR reactions were analysed on 2% agarose gels (Fig. 1a).

In-liquid spacer acquisition assay. Overnight cultures launched from single colonies were diluted 1:1,000 into a fresh 10-ml culture of BHI containing appropriate antibiotic and 5 mM $CaCl_2$. When the cultures reached $D_{600\,nm}$ of 0.4, depending on the experiment, they were either infected with phage MOI value of 1 (Fig. 1b) or induced with 1 μg ml^{-1} anhydrotetracycline (Fig. 1c). After 16 h, plasmids carrying the CRISPR systems were extracted using a slightly modified QIAprep Spin Miniprep Kit protocol: the pelleted bacterial cells were resuspended in 250 μl buffer P1 containing 50 ng μl^{-1} lysostaphin and incubated at 37 °C for 1 h, followed by the standard QIAprep protocol. 100 ng of plasmid DNA was used to amplify the CRISPR locus using Phusion DNA Polymerase (New England Biolabs) with the following primer mix: 3 parts JW8 and 1 part each of JW3, JW4 and JW5 (Extended Data Table 4). The following cycling conditions were used: (1) 98 °C for 30 s; (2) (for 30 times) 98 °C for 10 s, 64 °C for 20 s, 72 °C for 10 s; (3) 72 °C for 5 min. The PCR reactions were analysed on 2% agarose gels. To sequence individual spacers, the adapted bands were extracted, gel-purified and cloned via Zero Blunt TOPO PCR Cloning Kit (Invitrogen). CRISPR loci of individual clones were checked for expansion of the arrays by PCR using the primers listed above and sent for sequencing.

Phage adsorption assay. The phage adsorption assay was performed as described previously[30] with minor modifications. Cells were grown in BHI and 10 mM $CaCl_2$ to a $D_{600\,nm}$ (OD_{600}) of 0.4. The phage solution was prepared at 10^6 plaque-forming units (p.f.u.) per ml and 100 μl of this was added to 900 μl of cells. The mixture was incubated for 10 min at 37 °C to allow adsorption of the phage to the cellular membrane. The mixture was centrifuged for 1 min at 16,000*g* and the number of phage particles left in the supernatant was determined by phage titre assay.

Phage titre assay. Serial dilutions of the phage stock were prepared in triplicate and spotted on fresh top agar lawns of RN4220 in HI agar supplemented with the appropriate antibiotic and 5 mM $CaCl_2$. Plates were incubated at 37 °C overnight (Extended Data Fig. 2).

High-throughput sequencing. Plasmid DNA was extracted from adapted cultures using the in-liquid spacer acquisition assay described above. 100 ng of plasmid DNA was used as template for Phusion PCR to amplify the CRISPR locus with primers H182 and H183 (Extended Data Table 4). Following gel extraction and purification of the adapted bands, samples were subject to Illumina MiSeq sequencing.

Plasmid construction. Construction of pWJ40 was described elsewhere[17]. For the construction of pC194-derived and pE194-derived plasmids, cloning was performed using chemically competent *S. aureus* cells, as described previously[17]. The Δ*cas1* (pRH059), Δ*cas2* (pRH061) and Δ*csn2* (pRH063) mutants were constructed by one-piece Gibson assembly[31] from pWJ40 using the pairs of primers H016–H017, H018–H019, H020–H021, respectively (Extended Data Table 4). Plasmid pRH087 containing the wild type *cas* genes of *S. pyogenes* was obtained by inserting the first spacer of *S. pyogenes* (annealed primers H049 and H050 containing compatible BsaI overhangs) in pDB184 using BsaI cloning[32]. BsaI cloning was also used to construct pRH079 and pRH233 by inserting a ϕNM4γ4 targeting spacer (annealed primers H029 and H030) into pDB114 and pDB184, respectively. Plasmid pRH200 harbours the wild-type CRISPR3 system from *S. thermophilus* LMD-9 amplified with H168 and H169 from genomic DNA. The fragment was inserted on pE194 via Gibson assembly using H166 and H167. pRH213 was constructed by replacing $Cas9^{Sp}$ on pRH087 with $Cas9^{St}$ from pRH200 using the primer pairs H232–H233 and H231–H234, respectively. pRH214 was constructed by replacing $Cas9^{St}$ on pRH200 with $Cas9^{Sp}$ from pRH087 using the primer pairs H227–H230 and H228–H229, respectively. pGG32 was created by reducing the CRISPR locus of pWJ40 to a single repeat. This was accomplished by 'round the horn' PCR[33] using primers oGG82 and oGG83, followed by blunt ligation. pRH228 was constructed by replacing $Cas9^{Sp}$ on pGG32 with $Cas9^{St}$ from pRH200 using the primer pairs H232–H233 and H231–H234, respectively. pRH223 was constructed as a three-piece Gibson assembly combining TetR+ptet from pKL55-iTet (primers B534 and B616), pE194 (primers B532 and B617) and the *cas1*, *cas2*, *csn2* genes and the array from pGG32 (primers H176–H177). pRH231 was constructed from pGG32 by one piece Gibson assembly with primers H289–H290. pRH234 contains Cas1 E220A and was constructed via one-piece Gibson assembly from pRH223, respectively, using the primer pair H312–H313. pRH227 was constructed from pGG32 via two sequential single-piece Gibson assemblies: first, D10A was introduced with B337–B338 and second, H840A was introduced with B339–B340. pRH229 was constructed via one-piece Gibson assembly from pGG32 using the primer pair H276–H277. Plasmids pRH240, pRH241, pRH242, pRH243 and pRH244 were constructed by one-piece Gibson assembly with primers H237–H238 from pGG32, pRH228, pRH227, pRH229 and pRH231, respectively. pRH245 was constructed from pRH241 via two sequential single-piece Gibson assemblies: first, D10A was introduced with H336–H337 and second, H847A was introduced with H338–H339.

Isolation and sequencing of ϕNM4γ4. For the initial isolation of ϕNM4, supernatants from overnight cultures of *S. aureus* Newman were filtered and used to infect soft agar lawns of TB4:: ϕNM1,2 double lysogens[19]. A single plaque was picked and then plaque-purified in two additional rounds of infection using TB4 soft agar lawns, and subsequently used to lysogenize TB4. For the resultant lysogen, specific primers were used to verify the presence of ϕNM4 and the absence of ϕNM1,2 by colony PCR. High titre lysates of ϕNM4 (~10^{11} p.f.u. per ml) were then prepared from this lineage and used for infection of TB4/pGG9 soft agar lawns harbouring spacer 2B[17]. An escaper plaque was picked and then plaque-purified in two additional rounds of infection using TB4/pGG9 soft agar lawns. The resultant ϕNM4γ4 phage exhibited a clear plaque phenotype and was used to prepare a high titre lysate from which DNA was purified, deep sequenced, and assembled as described previously[17]. The full sequence of the ϕNM4γ4 has been deposited in GenBank under accession number KP209285, and includes a 2,784 bp deletion encompassing the C-terminal 80% of the φNM4 *cI*-like repressor gene.

Protein purification of Cas9. pMJ806 (wild-type Cas9) plasmid was obtained from Addgene. The proteins were purified as described before[6] with minor modifications as follows. The proteins were expressed in *E. coli* BL21 Rosetta 2(DE3) codon plus cells (EMD Millipore). Cultures (2 litres) were grown at 37 °C in Terrific Broth medium containing 50 μg ml^{-1} kanamycin and 34 μg ml^{-1} chloramphenicol until the D_{600nm} reached 0.6. The cultures were supplemented with 0.2 mM isopropyl-1-thio-β-D-galactopyranoside and incubation was continued for 16 h at 16 °C with constant shaking. The cells were collected by centrifugation and the pellets stored at −80 °C. All subsequent steps were performed at 4 °C. Thawed bacteria were resuspended in 30 ml of buffer A (50 mM Tris–HCl pH 7.5, 500 mM NaCl, 200 mM Li_2SO_4, 10% sucrose, 15 mM imidazole) supplemented with complete EDTA free protease inhibitor tablet (Roche). Triton X-100 and lysozyme were added to final concentrations of 0.1% and 0.1 mg ml^{-1}, respectively. After 30 min, the lysate was sonicated to reduce viscosity. Insoluble material was removed by centrifugation for 1 h at 16,200*g* in a Beckman JA-3050 rotor. The soluble extract was bound in batch to mixed for 1 h with 5 ml of Ni^{2+}-Nitrilotriacetic acid-agarose resin (Qiagen) that had been pre-equilibrated with buffer A. The resin was recovered by centrifugation, and then washed extensively with buffer A. The bound protein was eluted stepwise with aliquots of IMAC buffer (50 mM Tris-HCl pH 7.5, 250 mM NaCl, 10% glycerol) containing increasing concentrations of imidazole. The 200 mM imidazole elutes containing the His_6-MBP tagged Cas9 polypeptide was pooled together. The His_6-MBP affinity tag was removed by cleavage with TEV protease during overnight dialysis against 20 mM Tris-HCl pH 7.5, 150 mM KCl, 1 mM TCEP and 10% glycerol. The tagless Cas9 protein was separated from the fusion tag by using a 5 ml SP Sepharose HiTrap column (GE Life Sciences). The protein was further purified by size exclusion chromatography using a Superdex 200 10/300 GL in 20 mM Tris HCl pH 7.5, 150 mM KCl, 1 mM TCEP, and 5% glycerol. The elution peak from the size exclusion was aliquoted, frozen and kept at −80 °C.

Protein purification of Cas1. Plasmid pKW01 (wild-type Cas1) was constructed by through amplification of pWJ40 as a template for polymerase chain reactions (PCRs) to clone Cas1 into pET28b-His_{10}Smt3 using the primers PS192 and PS193 (Extended Data Table 4). Full sequencing of cloned DNA fragment confirmed perfect matches to the original sequence. The pKW01 plasmid was transformed into *E. coli* BL21 (DE3) Rosetta 2 cells (EMD Millipore). Cultures were grown and protein was purified by Ni-affinity chromatography step, as mentioned before in Cas9 purification. The 200 mM imidazole elutes containing the His_{10}-Smt3 tagged Cas1 polypeptide was pooled together. The His_{10}-Smt3 affinity tag was removed by cleavage with SUMO protease during overnight dialysis against 50 mM Tris-HCl pH 7.5, 250 mM NaCl, 20 mM imidazole and 10% glycerol. The tagless Cas1 protein was separated from the fusion tag by using a second Ni-NTA affinity step. The protein was further purified by size exclusion chromatography using a Superdex 200 10/300 GL in 20 mM Tris HCl pH 7.5, 500 mM KCl, 1 mM TCEP, and 5% glycerol. The elution peak from the size exclusion was aliquoted, frozen and kept at −80 °C.

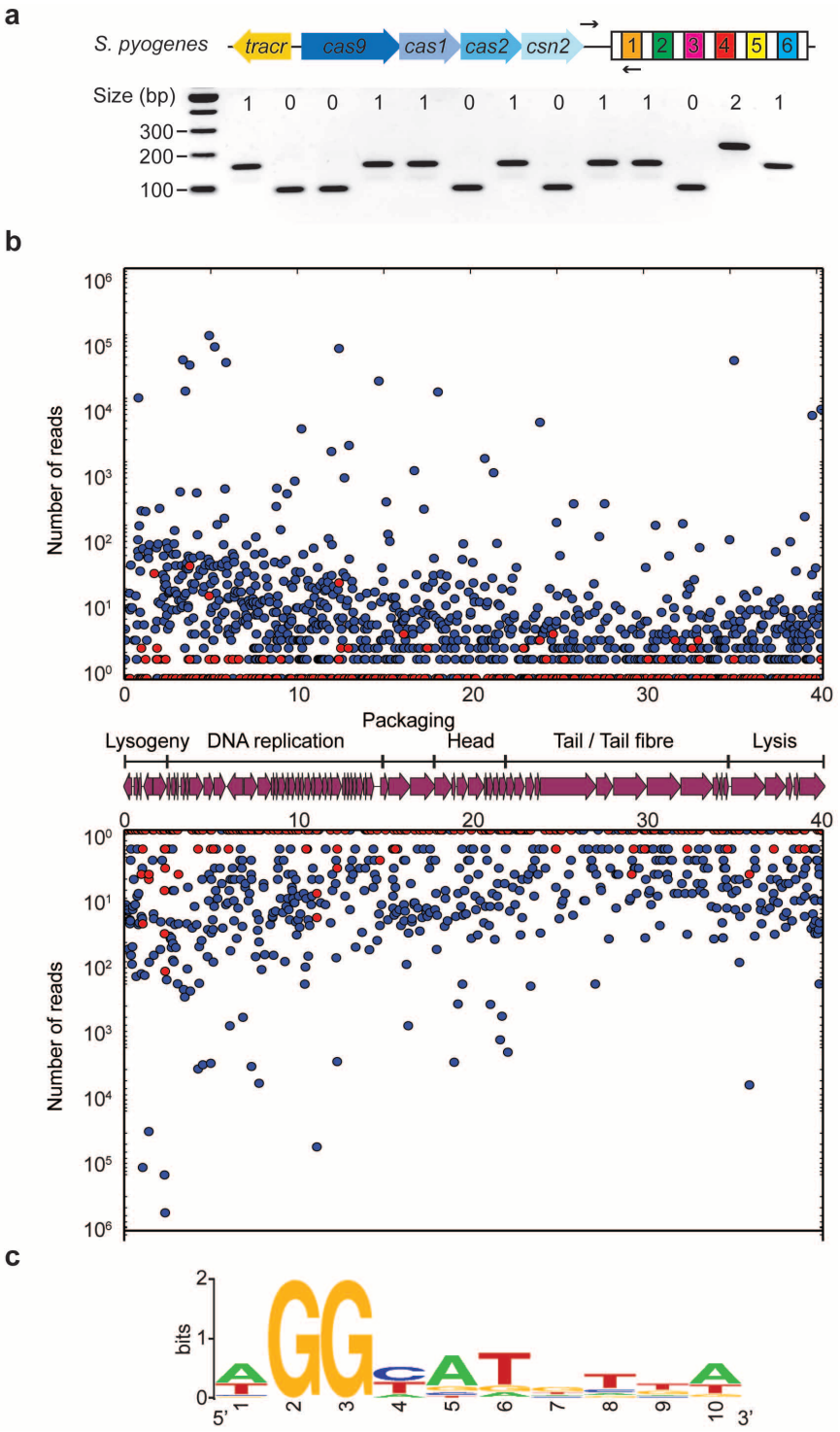
Protein purification of Cas2. The sequence encoding Cas2 was PCR amplified with primers PS334 and PS335 from pWJ40 and inserted into a pET-His_6 MBP TEV cloning vector (Addgene Plasmid number 29656) using ligation independent cloning (LIC). Sequencing of the resultant plasmid (pPS059) confirmed the matches to the wild-type sequence. The protein was expressed and purified following the same procedure as that for Cas9.

Protein purification of Csn2. Plasmid pPS060 was constructed by through amplification of pWJ40 as a template for polymerase chain reactions (PCRs) to clone Csn2 into pET28b-His_{10}Smt3 using the primers PS336 and PS337. Full sequencing of cloned DNA fragment confirmed perfect matches to the original sequence. Csn2 was expressed and purified following the same method as that of Cas1. Previously Csn2 was shown to form a tetramer[34]. Protein concentrations for all the purifications were determined by using the Bradford dye reagent with BSA as the standard.

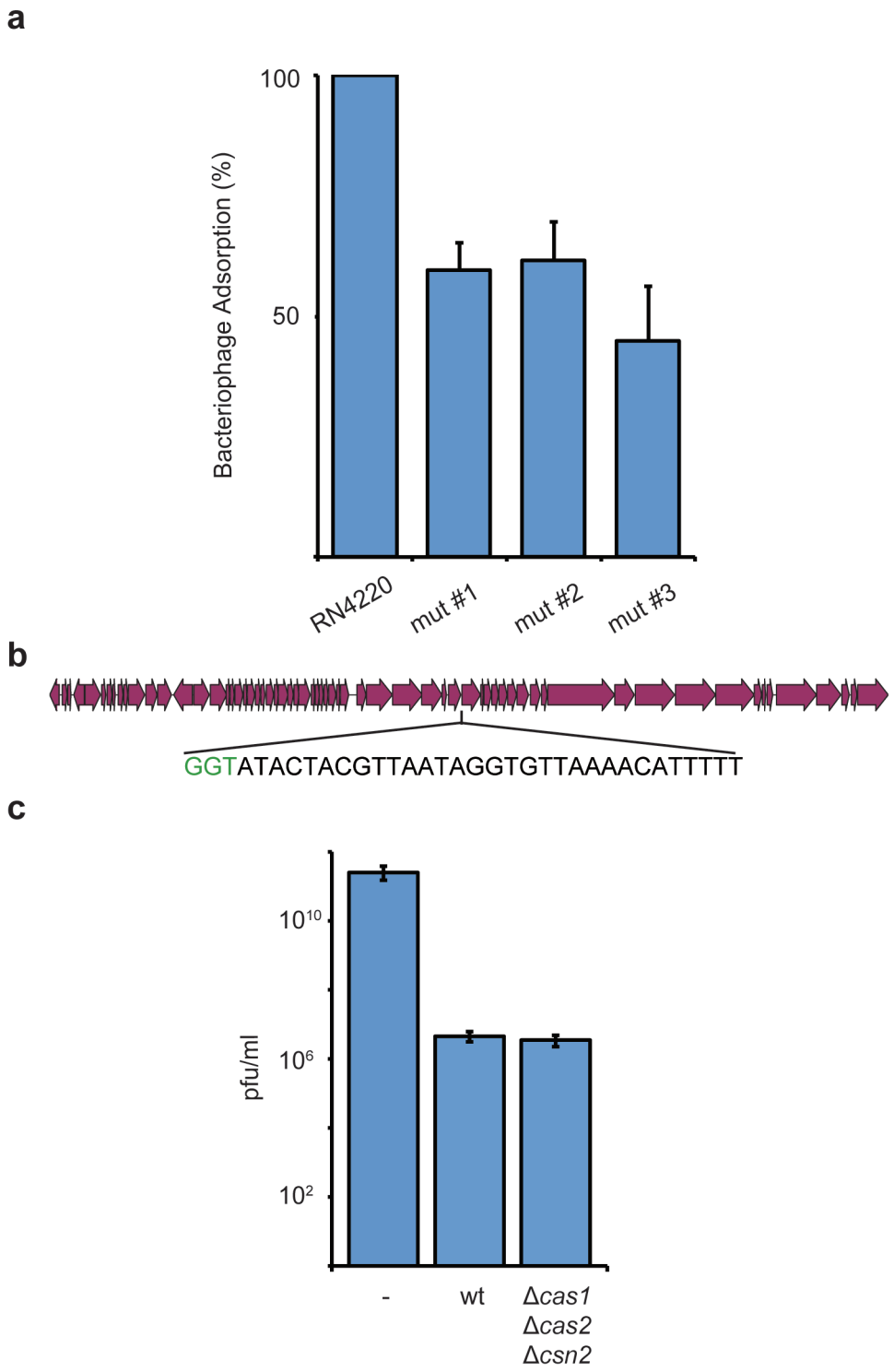
Protein purification of Cas9–Cas1–Cas2–Csn2 complex. pKW07 (His_{10}-Cas9–Cas1–Cas2–Csn2) was constructed by amplification of pWJ40 with primers PS199/PS202 and pET16b (Novagen) with primers PS200/PS203, followed by Gibson assembly of the fragments. Full sequencing of cloned DNA fragment was done to confirm perfect matches to the original sequence. The proteins were expressed in *E. coli* BL21 Rosetta 2(DE3) codon plus cells (EMD Millipore). Cultures were grown and protein was purified by Ni-affinity chromatography step, as mentioned before in Cas9 purification with minor modifications. The 200 mM imidazole eluates were dialysed overnight against 20 mM Tris-HCl pH 7.5, 150 mM KCl, 1 mM TCEP and 10% glycerol and subjected to mass spectrometry for the identification of the co-purifying proteins. pKW06 (Cas9–Cas1–Cas2–Csn2–His_6) was constructed by amplification of pWJ40 with primers PS204/PS205 and pET23a (Novagen) with primers PS206/PS207 (Extended Data Table 4), followed by Gibson assembly of the fragments. Full sequencing of cloned DNA fragment was done to confirm perfect matches to the original sequence. The proteins were expressed in *E. coli* BL21 Rosetta 2(DE3) codon plus cells (EMD Millipore). Cultures were grown and protein was purified by Ni-affinity chromatography step, as mentioned before in Cas9 purification with minor modifications. The 200 mM imidazole eluates were dialysed overnight against 20 mM Tris-HCl pH 7.5, 150 mM KCl, 1 mM TCEP and 10% glycerol. The proteins were further purified using a 5 ml SP Sepharose HiTrap column (GE Life Sciences), eluting with a linear gradient of 150 mM–1 M KCl.

Sample size. No statistical methods were used to predetermine sample size.

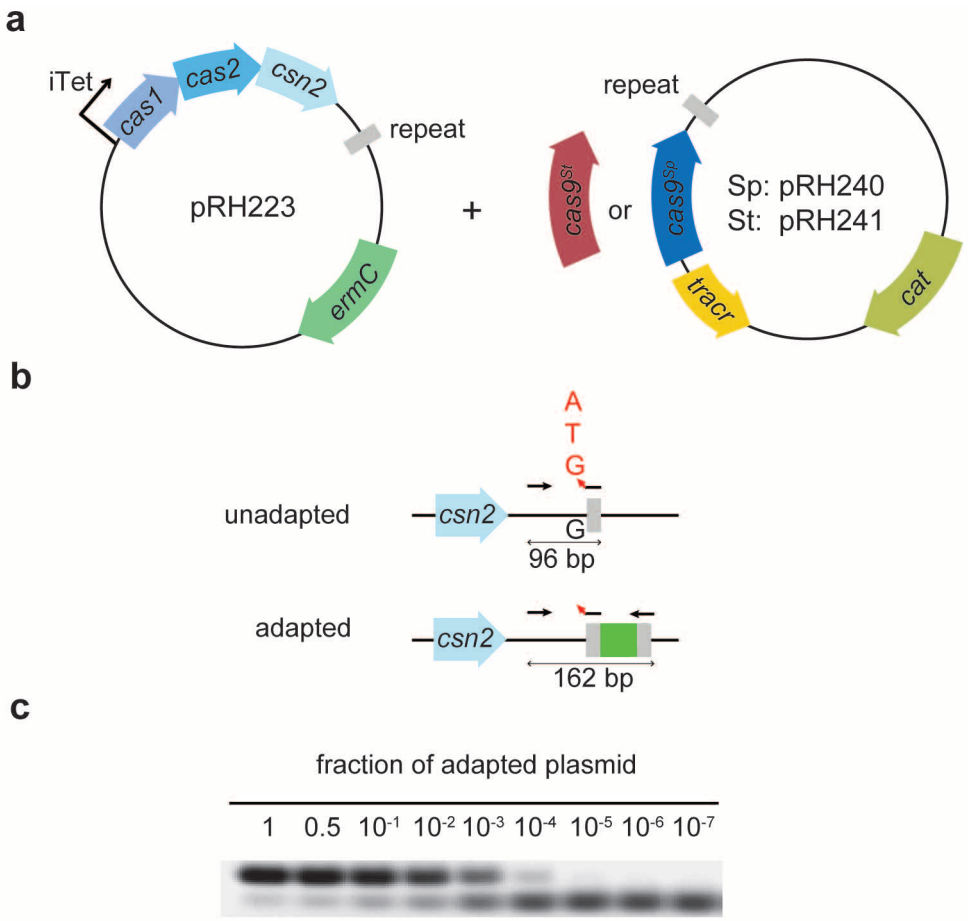
29. Horinouchi, S. & Weisblum, B. Nucleotide sequence and functional map of pE194, a plasmid that specifies inducible resistance to macrolide, lincosamide, and streptogramin type B antibodies. *J. Bacteriol.* **150,** 804–814 (1982).
30. Duplessis, M. & Moineau, S. Identification of a genetic determinant responsible for host specificity in *Streptococcus thermophilus* bacteriophages. *Mol. Microbiol.* **41,** 325–336 (2001).
31. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature Methods* **6,** 343–345 (2009).
32. Bikard, D. *et al.* Exploiting CRISPR–Cas nucleases to produce sequence-specific antimicrobials. *Nature Biotechnol.* **32,** 1146–1150 (2014).
33. Moore, S. D. & Prevelige, P. E., Jr. A P22 scaffold protein mutation increases the robustness of head assembly in the presence of excess portal protein. *J. Virol.* **76,** 10245–10255 (2002).
34. Arslan, Z. *et al.* Double-strand DNA end-binding and sliding of the toroidal CRISPR-associated protein Csn2. *Nucleic Acids Res.* **41,** 6347–6359 (2013).



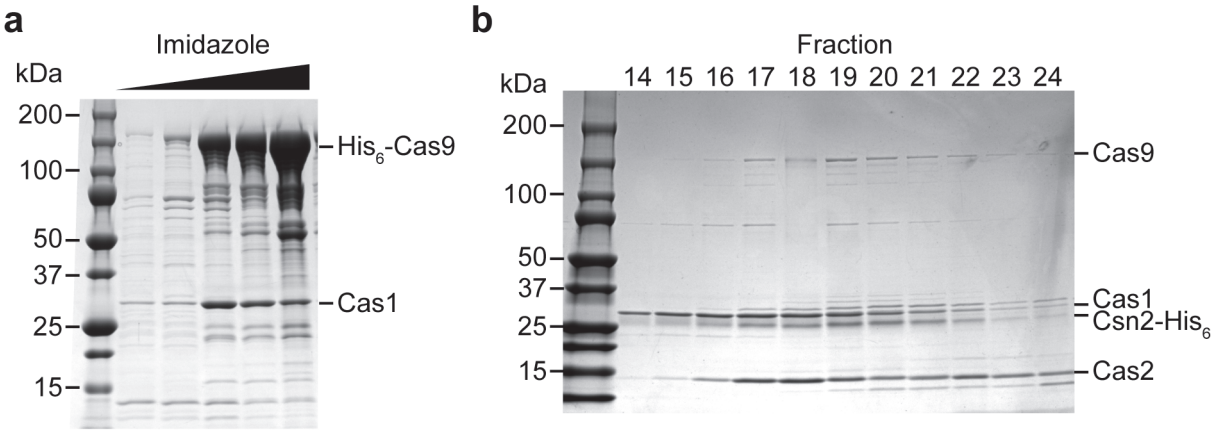
Extended Data Figure 1 | The *S. pyogenes* type II CRISPR–Cas system displays a strong bias for the acquisition of spacers matching viral protospacers with NGG PAMs. **a**, Analysis of bacteriophage-insensitive mutant colonies using PCR and agarose gel electrophoresis, representative of five technical replicates. Bacteria and phage were mixed in top agar and incubated overnight. DNA was isolated from individual colonies resistant to phage infection and used as template for a PCR reaction with primers (arrows) H182 and H183 (Extended Data Table 4), which amplify the 5′ end of the *S. pyogenes* CRISPR array. The size of the PCR band indicates the number of new spacers (shown at the top of the gel). Cells without additional spacers resist infection by a CRISPR-independent mechanisms, presumably envelope resistance. **b**, Analysis of acquired spacers during phage infection of a population of bacteria carrying the *S. pyogenes* type II CRISPR–Cas system. Liquid cultures of bacteria were infected with phage, surviving cells were collected at the end of the infection, DNA extracted and used as template for a PCR reaction as described above. Amplification products were separated by agarose gel electrophoresis and the DNA of the bands corresponding to products with additional spacers was extracted and sent for Mi-Seq next-generation sequencing. Reads corresponding to newly acquired spacers were plotted according to their position in the phage ϕNM4γ4 genome (*x* axis) and their abundance (*y* axis). Each dot represents a unique spacer sequence; blue and red dots indicate a corresponding protospacer with an NGG or non-NGG PAM. Top and bottom plots indicate protospacers in the top and bottom strands of the ϕNM4γ4 DNA. The map as well as the different functions of the phage genes are indicated in between the plots. The raw data used to make this graph is in the Source Data File . **c**, Weblogo showing the conservation of the 5′ flanking sequences of 10,000 protospacers randomly selected from the experiment shown in **b**. Absolute conservation of the NGG PAM was observed.



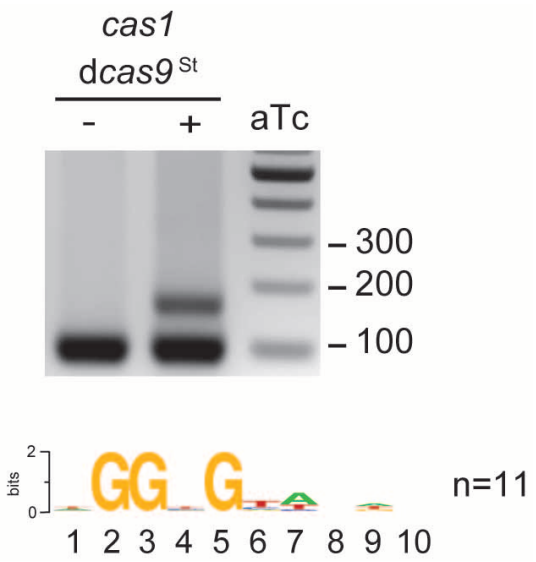
Extended Data Figure 2 | *cas1*, *cas2* and *csn2* are not required for the execution of immunity. **a**, Analysis of bacteriophage-resistant mutants that do not acquire a new spacer. Three colonies that survived phage infection in our in-plate adaptation assay (Extended Data Fig. 1) were subjected to phage adsorption assay. Briefly, surviving colonies as well as the wild-type *S. aureus* RN4220 control were grown in liquid and mixed with bacteriophage. After a brief incubation, cells were pelleted by centrifugation and the phages present in the supernatant (unable to bind and infect cells) were counted on a lawn of sensitive cells. The number of plaque-forming units (p.f.u.) of a control experiment in the absence of host cells were used to determine the 100% free-phage, or 0% adsorption value. No plaques were observed in the control experiment using wild-type cells and this value was used to set the 100% adsorption limit. The three CRISPR-independent, bacteriophage-resistant mutants displayed a marked defect in phage adsorption (about 50%), indicating that most likely they carry envelope resistance mutations. Error bars: mean ± s.d. ($n = 3$). **b**, *cas1*, *cas2* and *csn2* are not required for the execution of immunity using previously acquired spacers. Position within the phage ϕNM4γ4 genome of the type II CRISPR–Cas target used in this experiment. The protospacer sequence is in the bottom strand (shown in 3′–5′ direction) and flanked by a TGG PAM (in green). **c**, Comparison of immunity provided by a type II CRISPR–Cas system programmed to target the sequence shown in panel **a** in the presence (wild-type, wt) or absence (Δ*cas1*,Δ*cas2*, Δ*csn2*) of *cas1*, *cas2* and *csn2*. Immunity is measured as the p.f.u. of a ϕNM4γ4 phage lysate spotted on top agar lawns of *S. aureus* RN4220 cells containing no CRISPR system (−), a wild-type *S. pyogenes* CRISPR–Cas type II system (wt, pRH233), or the same CRISPR–Cas systems with a deletion of *cas1*, *cas2* and *csn2* genes (Δ*cas1*, Δ*cas2*, Δ*csn2*, pRH079). Error bars: mean ± s.d. ($n = 3$).



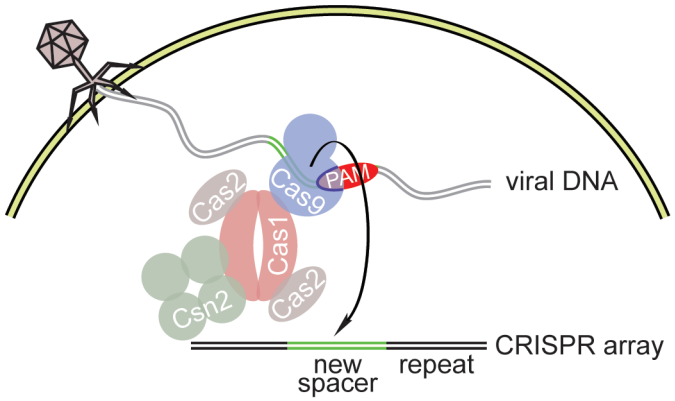
Extended Data Figure 3 | Generation of an experimental system for the overexpression of *cas1*, *cas2* and *csn2* and the detection of spacer acquisition in the absence of phage infection. **a**, Plasmids used in the spacer acquisition experiments presented in Figs 1c and 2c, d. pRH223 contains *cas1*, *cas2* and *csn2* from *S. pyogenes* under a tetracycline-inducible promoter. Cells containing this plasmid only acquired spacers when a second plasmid expressing *cas9* was introduced, pRH240 or pRH241, containing the tracrRNA gene, the leader and first repeat from the *S. pyogenes* type II CRISPR–Cas system as well as *cas9* from *S. pyogenes* (*cas9*^{Sp}) or *S. thermophilus* (*cas9*St), respectively. The leader is a short, AT-rich sequence immediately upstream of the first repeat that contains the promoter for the transcription of the CRISPR array. **b**, Highly sensitive PCR assay to enrich for amplification products of adapted CRISPR loci. Arrows indicate primer annealing position and direction. The forward primer (JW8) anneals on the leader. For the reverse primer, a cocktail of JW3, JW4 and JW5 was used. The three reverse primers anneal on the repeat and differ only in their 3′-end nucleotide that never matches the last nucleotide of the leader (red arrowhead). Because this nucleotide is critical for the annealing of the primers, loci that acquire spacers ending in A, C or T are preferentially amplified over unadapted loci. **c**, To quantify the sensitivity of this technique, we mixed pGG32 (one repeat, unadapted) with pRH087 (repeat-spacer-repeat, adapted) in known ratios. The amplification of adapted plasmid was detected even when it represented 0.01% (10^{-4}) of the total plasmid template, representative of three technical replicates. This highly sensitive PCR assay is not required to detect acquisition during phage infection, as in this case adapted cells survive and are enriched within the population, making their detection much easier.



Extended Data Figure 4 | Purification of a Cas9–Cas1–Cas2–Csn2 complexes. **a**, The *cas9–cas1–cas2–csn2* operon of *S. pyogenes* SF370 was cloned into the pET16b vector (generating pKW07) to add an N-terminal histidyl tag to Cas9 and express all proteins in *E. coli*. Purification was performed using Ni-NTA affinity chromatography. SDS–PAGE followed by Coomassie staining of the purified proteins revealed a co-purifying protein that was identified as Cas1 by mass spectrometry, in a result representative of five technical replicates. Mass spectrometry identification of all the eluted proteins co-purifying with Cas9 is shown in Extended Data Table 2. **b**, The *cas9–cas1–cas2–csn2* operon of *S. pyogenes* SF370 was cloned into the pET23a vector (generating pKW06) to add an C-terminal histidyl tag to Csn2 and express all proteins in *E. coli*. Purification was performed using Ni-NTA affinity chromatography followed by ion exchange chromatography. The elution fractions that constituted the peak containing the complex (Fig. 3a) were separated by SDS–PAGE and visualized by Coomassie staining, representative of three technical replicates.



Extended Data Figure 5 | dCas9St can also support spacer acquisition. A plasmid derived from pRH241 containing mutations in the active site of *S. thermophilus* Cas9 (D10A, H847A; dCas9St) was used to characterize spacer acquisition in the absence of phage infection. Upon overexpression of Cas1, Cas2 and Csn2 using anydrotetracycline (aTc), we were able to detect spacer acquisition. Sequencing of spacers and alignment of the protospacer flanking sequences demonstrated the selection of an NGGNG PAM. The image is representative of three technical replicates.



Extended Data Figure 6 | A model for the selection of PAM-flanking spacers by Cas9. After injection of the phage DNA, an adaptation complex formed by Cas9, Cas1, Cas2 and Csn2 uses the Cas9 PAM binding domain to specify functional protospacers, that is, that are followed by the correct PAM. It is not known how the protospacer sequence is extracted from the viral DNA to become a spacer. In the 'cut and paste' model, a nuclease, possibly Cas1, cuts the viral DNA to generate the spacer. In the 'copy and paste' model the protospacer sequence is copied first. Once loaded with the selected protospacer sequence, this complex promotes the integration of this sequence into the CRISPR array, thus becoming a new spacer. Previous studies demonstrated that Cas1 dimerizes and interacts with Cas2 (ref. 13); Csn2 has been determined to form a tetramer[34].

Extended Data Table 1 | Sequences of the spacers analysed to obtain the sequence logos in this study

Figure	Spacer	Sequence	PAM	Target
2b	1	gcaacaatgggaaccaagctatgttgatag	aGGgt	phage
1st logo	2	gagaacaaaaccatcctacccggtaataaa	tGGta	phage
	3	aatagagatactttatctaacatgatacac	gGGag	phage
	4	ccatttagatttcaaaagtttagtatctat	aGGca	phage
	5	agtattggaatctgatgaatattcatctct	cGGta	phage
	6	agaaaatttatacattgattattcaccaac	aGGca	phage
	7	acatactccaaacaattgatggatttgtgt	aGGtg	phage
	8	gctaagactgtgaagcataatactgctact	aGGta	phage
	9	ttttaagctattcattttaaaaggtcatat	gGGca	phage
	10	acttatgccgtttctatacttcactacagca	tGGtc	phage
	11	atgaatggattgaagagaacacagacgaac	aGGac	phage
	12	ccacaaatagaaatagagctagggagtttaa	cGGta	phage
	13	attagttactccacaaatagaaatagagct	aGGga	phage
	14	ggagtaactaatatctgaattgttatcagt	tGGtt	phage
	15	tagtttttgagtatgcttacttttttcttg	tGGtt	phage
	16	tgaacgaattgtcagtatgtacagattaat	aGGaa	phage
	17	cattacggacgtagtagaagcaattagaaa	tGGaa	phage
	18	tggatatgacgaccaagatttagcgtttta	aGGtg	phage
	19	cgacataacgctaatacatgtttgtcatag	tGGtt	phage
	20	acaaacttaacaatagtggttttttcaaga	gGGag	phage
2b	1	agagtacaatattgtcctcattggagacac	tGGgG	phage
2nd logo	2	tgtttgggaaaccgcagtagccatgattaa	gGGtG	phage
	3	ctcatattcgttagttgcttttgtcataaa	aGGtG	phage
	4	tttatgtctatatactcaaagtaatcatttt	cGGaG	phage
	5	taatatcaacggtatgtgggtgtctggtga	cGGtG	phage
	6	aataagtctaaaaaaccaacgtttaatgat	tGGgG	phage
	7	gttgatattacgttcatagaacatacctga	tGGtG	phage
	8	tcaatgtttggtacaagttggtcacagata	tGGaG	phage
	9	ttagttactccacaaatagaaatagagcta	gGGaG	phage
	10	caattgtttttcttggaaatcatatttata	cGGcG	phage
	11	tatctaagtttgccaattattacattaaagc	tGGtG	phage
	12	taggacatagagatgaaaaaacgactataa	aGGtG	phage
	13	tgaagaaatgattcaagaaacacaaaagag	tGGcG	phage
	14	tcggactgttagggtacgcgaagggcaaaa	aGGaG	phage
	15	aatactttcttctaaaaaacctaagtcaac	aGGaG	phage
	16	taatccaattacaacattaaaaattaatga	cGGaG	phage
	17	acaatgttaagcaaccagcacattacacata	cGGcG	phage
	18	ggattttaaaataaaagtaaatgttgatac	tGGcG	phage
	19	caggcaatgttatttttatcggatttttaaaaa	cGGcG	phage
	20	agaatctttattattagctgacttacaaga	aGGtG	phage
	21	aaaaccccaatatcttttaaaaataaagtt	aGGtG	phage
	22	tagggcaatgattgaagaatttgatgataa	cGGaG	phage
2b	1	aaaggcaacatatttgaatcatcacatttat	tGGaG	phage
3rd logo	2	ttggaatggaattaaacaataaaactttta	tGGaG	phage
	3	atattcatcagattccaatactacgttaat	aGGtG	phage
	4	acaatttaaaaattagaaatgtaaatgtag	aGGtG	phage
	5	cagaatgaactatgaaacaggggtccaact	aGGtG	phage
	6	acataacatcaaaaccctttctgaagaaat	tGGtG	phage
	7	taagttgtttgaaatgtacgagatggaagg	aGGaG	phage
	8	atacgtgtaaagacatattagatcgagtca	aGGaG	phage
	9	tgtgcaggagctacgttcaataaatgtgaa	aGGaG	phage
	10	ttaagaaagttattgtcatcgagcttaaat	tGGtG	phage
	11	acacacatactaaacctgaacgattaagga	gGGgG	phage
	12	tttaccaacatccttagttgatagattttt	aGGcG	phage
	13	gtttgaatacgttccgtttctgatacccagt	aGGcG	phage
	14	aagttaaaaagaatttaaagtcaagaagta	tGGgG	phage
	15	attctcagaagatagcgaagatgggagaaa	aGGaG	phage
	16	tgagcgactgctgggtgtgcttcgaatagtt	tGGcG	phage
	17	taatatatgctcatacttaattgaattgtc	tGGtG	phage
	18	atcttctttttttaatacgtccatcaacaag	cGGtG	phage
	19	cgatattggcggtgtgaataataactttaa	aGGaG	phage
	20	caacgagctggcaacaacataagatgacag	aGGcG	phage
2b	1	taaactactacgacttaagcaggtgccata	tGGca	phage
4th logo	2	gacaaatgctattcaacattcagttaaaga	aGGta	phage
	3	acaattattaattgaacaagcgcaagctaa	cGGct	phage
	4	cacatcaattagtaagacgccaaaagtaac	aGGta	phage
	5	aaacgatgagtacacaaaatacaaaatcta	cGGca	phage
	6	gtaataatatttttaataacctcaacatct	tGGtc	phage
	7	tcatgaaaaagtgaattgctagtagtgtgt	tGGtc	phage
	8	tacgctatcgcaaaagcagtcaaagctaaa	gGGca	phage
	9	agggaatctacagttattaaataactatt	tGGat	phage
	10	aaaacgagcaaattaagtggtacgtagaca	aGGgt	phage
	11	ctaaatgttgccatttcgttatctcctttc	tGGta	phage
	12	actggatgacattgaacaaagcaccgaata	tGGcc	phage
	13	taaatatttgataacaacattatacacgaa	aGGag	phage
	14	cacatcaattagtaagacgccaaaagtaac	aGGta	phage
	15	aaggtgatgacggcgaatggtacacaacata	tGGtc	phage
	16	taacgacggtacttattccgtcgttgctac	tGGtg	phage
	17	ataaataaaaaagttactactcacacacta	aGGca	phage
	18	tctaggttcgaactcttctttaaatttaat	aGGca	phage
	19	ctcatcaatatcattctgattggttatttt	gGGat	phage
	20	tctctttgataaataactttatccacataa	aGGtg	phage
	21	ttagacttttacttttccattacttaaatca	tGGtc	phage
	22	aatttgttcttgcgcttcaatagtgatagt	aGGgt	phage
	23	ataagtctaaaaaaccaacgtttaatgatt	gGGga	phage
2d	1	acatgttatgcatatcgtaagtgaagtcac	aGGta	chromosome
1st logo	2	agatcaaattgtaacaactaatcctattgc	aGGta	chromosome
	3	gtttcagcaatatatctcttagtgcatcac	cGGtt	chromosome
2d	4	tacaatgtaggctgctctacacctagcttc	tGGgc	pRH223
1st logo	5	tttgattacaatggcacatgtacttatgcc	tGGat	chromosome
	6	catttgtcttagcacatgaattaggtcatgc	aGGtc	chromosome
	7	catgattgcacccattgttgcacctagtac	aGGtt	chromosome
	8	taccaataacttaagggtaactagcctcgc	cGGca	chromosome
	9	gagtatgtttgcgcgtgaagtggttgtgtc	tGGat	pRH223
	10	agaatggttagatttatggcgtgatgtaac	gGGca	chromosome
	11	ctgcttccatgataactggaccatcagcaac	cGGat	chromosome
	12	gattgaagctacaatacctgatgttgctgc	gGGaa	chromosome
	13	cgaaatacttggctaagcacgacgaggcct	tGGtg	pRH223/240
	14	gctctacacctagcttctgggcgagtttac	gGGtt	pRH223
	15	atatggaagttacatttttttggaacgagtgc	aGGtt	chromosome
	16	ttatgaagcgttacgtcaacaagattttcc	aGGat	chromosome
	17	ttatcgaagtatacgagttcacagaagaac	aGGct	chromosome
	18	ccagttcttgttgttttggtgctttagtca	aGGtt	chromosome
	19	tggatgatcttgtctttcatgtgtacctgt	tGGaa	chromosome
	20	caggatttagtttttcctagcggtcatgcta	tGGga	chromosome
2d	1	ttgactatcaaatgtctttttcaatgtttc	gGGtG	chromosome
2nd logo	2	atccgttctgcagaagagattgtttcttgc	aGGcG	pRH223
	3	tgaacatttcgattatgtattaatgagtgc	tGGtG	chromosome
	4	catctttaggacgaatgccagcacgttctgc	tGGaG	chromosome
	5	caccatgttaaaaatacctccatcatcacc	aGGaG	pRH223/241
	6	tcgtgagacagttcggtccctatccgtcgt	gGGcG	chromosome
	7	tttgcgcagtcggcttaaaccagttttcgc	tGGtG	pRH223
	8	aaagaagtcataagtaccatgacttgagtt	tGGtG	chromosome
	9	ctaatttttcttcttcaacaccatctatggc	tGGcG	chromosome
	10	ccaagtattcaaagttggaacgggtggtct	aGGtG	chromosome
	11	atccgttctgcagaagagattgtttcttgc	aGGcG	pRH223
	12	tttgcgcagtcggcttaaaccagttttcgc	tGGtG	pRH223
	13	aacgcgtatacatagcaagcgttctcatgt	tGGaG	chromosome
	14	agtttgggagtcaattatcggctttttaac	tGGcG	chromosome
3d	1	tgacttctctgaagagccatctttttgcact	tGGaa	chromosome
1st logo	2	ggtcagatgcaattcgacatgtggacggac	tGGtt	pRH223
	3	atcttttctagctttttctccaagcacagac	aGGac	chromosome
	4	gttggtctaattgtttcaatagttccacct	tGGtc	chromosome
	5	tgccggttggggtggctgagacggcaccct	aGGaa	chromosome
	6	tgagtatgtttgcgcgtgaagtggttgtgtc	tGGat	pRH223
	7	ttgagttagaaaacggtcgtaaacggatgc	tGGct	pRH242
	8	agtttgggagtcaattatcggctttttaac	tGGcg	chromosome
	9	aattaagaaatcttctaaccaactgattgc	tGGaa	chromosome
	10	aacagaaagaatagggaaggtatccgactgc	tGGta	pRH223/242
	11	tggtattgtaggcgttattttaggtattcc	gGGat	chromosome
	12	aaatctcagcaggacaagctggtacaggtgc	tGGtt	chromosome
	13	ctcaagagatttggagcatccaatcaatgc	aGGtc	pRH223
	14	ctaaggtggcaccacggtaacgcgtccttac	aGGta	chromosome
	15	tgattaaacttaaaaatgtattacctagtgc	aGGta	chromosome
	16	atttgagtcagctaggaggtgactgatggc	tGGtt	pRH223
	17	ataagagaagatgctagacgtataagttcac	tGGtc	chromosome
	18	acgttttatctgtatttgcgacaatcgttg	gGGta	chromosome
	19	ataacatacgccgagttatcacataaaagc	gGGaa	pRH223
	20	gcattttaaacaaaaaaagatagacagcac	tGGca	pRH223
	21	aatcccagttagaacaaacgctaaaatggc	gGGcc	chromosome
	22	taccaataacttaagggtaactagcctcgc	cGGca	chromosome
3d	1	gaagtctagctgagacaaatagtgcgatta	caaaa	pRH223/243
2nd logo	2	agcatagctctaaaacctcgtagactattt	ttgtc	pRH223/243
	3	aaatttttagacaaaaatagtctacgag	gtttt	pRH223/243
	4	aagtcgaacttcataatcatcgctttcgg	catat	chromosome
	5	ccaatttctacagacaatgcaagttggggt	gtggg	chromosome
	6	gttatttctgaaatgcccgttacatcacgc	cataa	pRH243
	7	tgtttgccctccaaatatgaaaacatggcc	cggta	chromosome
	8	atgagatgaggcgataaaagaacgtcgcta	aaacg	chromosome
	9	tactacttcaaggaattctatagaacctac	tatat	chromosome
	10	gtaccacagtgccacatgttggcaattggc	gagac	chromosome
	11	taaagctggtgaagcgattaacactgtacc	aagta	chromosome
	12	atttcttcgttattagaaaatataaaattgc	gttgt	chromosome
	13	atttttttatgattaagccatatggggttaa	gcaag	pRH223/243
	14	aaagactgggatccaaaaaaatatggtggt	tttga	pRH243
	15	attttcaaatgcataaaaactgtttctcaac	gatat	chromosome
	16	ttttgtattggaatggcatttttgctatc	aaggt	chromosome
	17	taaaacaggaccacttgtcatgtaagcttt	aagtt	pRH223/243
	18	tgtatcttgtggtttcatctgtgctaacttt	ggcag	chromosome
	19	agaggatgcagaacgtgcaatcttagctgc	aagac	chromosome
	20	ttcaaacgagaataattatggcgttggttta	ggtat	chromosome
	21	gcgaatacactcattaaaacaattgcatcc	tgatt	chromosome
	22	tcttatcttgataataagggtaactattgc	cgatg	chromosome
	23	caaataaaggtgcgttattaataacagtgc	caggc	chromosome
	24	acaacagtacgccaaccagccatcagtcac	ctcct	pRH223
Ext. Data	1	cagctaacaatgccatgattggtcggctga	gGGaG	pRH223
Fig. 5	2	tggtaaatttacagaagatgctgaagatgc	tGGtG	chromosome
	3	gagtcagctaggaggtgactgatggctggt	tGGcG	pRH223
	4	caaataagtctagacatattagctcgttatc	aGGtG	chromosome
	5	acgaccttgttgcaacatagcgccccactc	tGGtG	chromosome
	6	acatgttatgcatatcgtaagtgaagtcac	aGGtG	chromosome
	7	atccgttctgcagaagagattgtttcttgc	aGGcG	pRH223
	8	agatgcttgttgtgttgtttgtgttgatgc	cGGtG	chromosome
	9	atccgttctgcagaagagattgtttcttgc	aGGcG	pRH223
	10	agtttgggagtcaattatcggctttttaac	tGGcG	chromosome
	11	aaaaagttatctcgtagacattacactggc	tGGgG	pRH245

Extended Data Table 2 | Mass spectrometry analysis of proteins purified through Ni-NTA

Accession	Protein	% Coverage	Unique Peptides	Total peak area
	Cas9	83.26	170	9.7×10^{10}
	Cas1	91.35	40	1.9×10^{10}
	Cas2	84.07	13	1.9×10^{9}
	Csn2	91.82	18	2.9×10^{9}
P77398	Bifunctional polymyxin resistance protein ArnA (*arnA*)	85.76	43	8.2×10^{8}
P60422	50S ribosomal protein L2 (*rplB*)	67.40	24	1.9×10^{9}
P17169	Glucosamine--fructose-6-phosphate aminotransferase (*glmS*)	79.31	38	1.8×10^{8}
P0AA43	Ribosomal small subunit pseudouridine synthase A (*rsuA*)	85.71	17	8.9×10^{8}
P0A9K9	FKBP-type peptidyl-prolyl cis-trans isomerase (*slyD*)	68.88	7	3.7×10^{9}
P0ACJ8	Catabolite gene activator (*crp*)	82.86	18	5.4×10^{8}
P45395	Arabinose 5-phosphate isomerase (*kdsD*)	73.17	21	1.2×10^{8}
P0A6F5	60 kDa chaperonin (*groL*)	83.94	38	2.8×10^{8}
P0A9A9	Ferric uptake regulation protein (*fur*)	78.38	8	1.2×10^{9}
P08622	Chaperone protein DnaJ (*dnaJ*)	72.07	19	1.4×10^{9}
P00393	NADH dehydrogenase (*ndh*)	59.22	16	3.6×10^{8}

Extended Data Table 3 | Mass spectrometry analysis of protein bands from the purified Cas9–Cas1–Cas2–Csn2 complex

Protein	% Coverage	Unique Peptides	Total peak area
Cas1	67.82	26	3.4×10^{8}
Cas2	90.27	13	1.2×10^{9}
Cas9	68.49	111	4.1×10^{8}
Csn2	82.27	19	4.1×10^{8}

Extended Data Table 4 | Oligonucleotides used in this study

Primer	Sequence
B337	gacgctatttgtgccgatagctaagcctattgagtatttc
B338	gaaatactcaataggcttagctatcggcacaaatagcgtc
B339	ggaaactttgtggaacaatggcatcgacatcataatcact
B340	agtgattatgatgtcgatgccattgttccacaaagtttcc
B532	ctttttccgtgatggtaactgttcatatttatcagagctcgtg
B534	gagctctgataaatatgaacagttaccatcacggaaaaaggttatg
B616	ttattttaattatgctctatcaa
B617	gagtgatcgttaaatttatactgc
H016	aggaggtgactgatgggagttcctgaatttaggatatgag
H017	taaattcaggaactcccatcagtcacctcctagctgactc
H018	ttaggatatgagtgaggcttttgatgaatcttaatttttc
H019	ttcatcaaaagcctcactcatatcctaaattcaggaactc
H020	tttgatgaatcttaataaaaatatggtataatactcttaa
H021	ttataccatatttttattaagattcatcaaaagcctcccc
H029	aaacaaaaatgttttaacacctattaacgtagtatg
H030	aaaacatactacgttaataggtgttaaaacattttt
H049	aaactgcgctggttgatttcttcttgcgctttttg
H050	aaaacaaaaagcgcaagaagaaatcaaccagcgca
H166	gaaatgtgagaagggacctctgataaatatgaacatgatgagtgatcg
H167	ggactcttttatctctactcgtgctataattatactaattttataaggagg
H168	agtataattatagcacgagtagagataaaagagtcctttggatgattcc
H169	tgttcatatttatcagaggtcccttctcacatttcaatactagactc
H176	ttgatagagcataattaaaataagatgccactcttatccatcaatcc
H177	gcagtataaatttaacgatcactctaaaacctctccaactacctccc
H182	nnnnncagcaaaatttttagacaaaaatagtc
H183	nnnnncagaagaagaaatcaaccagcgc
H227	taatggcaggttggagaacagtagtc
H228	actactgttctccaacctgccattagtcacctcctagctgactc
H229	agatttttcaaataaggagaaatgtttgaaatcatcaaactcattatggatttaatttaaacttttttattttagg
H230	acatttctccttatttgaaaaatctaaatttatagaaattattatacgc
H231	aacttttttattttaggaggcaaaaagcgtataataatttctataaatttagatttttcaaataagg
H232	ttttgcctcctaaaataaaaagtttaaattaaatccataatgag
H233	tgatggctggttggcgtac
H234	caacagtacgccaaccagccatcaaccctctcctagtttggc
H237	ggcgtactgatgaagattatttcttaataactaaaaatatgg
H238	tttagttattaagaaataatcttcatcagtacgccaaccagcc
H276	ttgatcaaaaacaatatacgtctacaaaagaag
H277	tagacgtatattgtttttgatcaattgttgtatcaa
H289	agcgcttgggagaaattcaaagaaatttatcagcc
H290	tttctttgaatttctcccaagcgctttcaaaacgc
H312	gatattatggcaccatttaggcctttagtgg
H313	aaaggcctaaatggtgccataatatcgctagc
H336	catactcaattggacttgctattggaacgaatagtgttgg
H337	cgttccaatagcaagtccaattgagtatggcttagtc
H338	gtaattatgatattgatgctattattcctcaagc
H339	gaggaataatagcatcaatatcataattacttaatc
JW3	aaaacagcatagctctaaaacg
JW4	aaaacagcatagctctaaaaca
JW5	aaaacagcatagctctaaaact
JW8	ggcttttcaagactgaagtctag
L400	cgaaatttttagacaaaaatagtc
oGG82	aacattgccgatgataacttgag
oGG83	gttttgggaccattcaaaacagcatagctctaaaacctcgtag
PS192	CGCGGATCCATGGCTGGTTGGCGTACTGTTGTGG
PS193	CGCCTCGAGTCATATCCTAAATTCAGGAACTCC
PS199	CGAGCATATGACGACCTTCGATATGATCGGCAATGTTGAATGGAGACCATTC
PS200	GAATGGTCTCCATTCAACATTGCCGATCATATCGAAGGTCGTCATATGCTCG
PS202	CATCATCATCATCATCACAGCAGCGGCATGGATAAGAAATACTCAATAGG
PS203	CCTATTGAGTATTTCTTATCCATGCCGCTGCTGTGATGATGATGATGATG
PS204	CGACAAGCTTGCGGCCGCACTCGAGCTTTTTATTTTAGGAGGCAAAAATG
PS205	GGATCTCAGTGGTGGTGGTGGTGGTGTACCATATTTTTAGTTATTAAGAAATAATC
PS206	GATTATTTCTTAATAACTAAAAATATGGTACACCACCACCACCACCACTGAGATCC
PS207	CATTTTTGCCTCCTAAAATAAAAAGCTCGAGTGCGGCCGCAAGCTTGTCG
PS284	GCTAGCGATATTATGGcACCATTTAGGCCTTTAG
PS285	CTAAAGGCCTAAATGGTgCCATAATATCGCTAGC
PS334	TACTTCCAATCCAATGCAATGAGCTATCGCTATATG
PS335	TTATCCACTTCCAATGTTATTATTAGCTTTCATCAAAGGC
PS336	CGCGGATCCATGAACCTGAACTTTAGCCTGCTGG
PS337	CGCCTCGAGTTACACCATATTTTTGGTAATCAG
PS354	GTTCCTGAATTTAGGATATGAAACATTGCCGATCATATCGAAGG
PS355	CCTTCGATATGATCGGCAATGTTTCATATCCTAAATTCAGGAAC

LETTER

doi:10.1038/nature14167

Regulation of star formation in giant galaxies by precipitation, feedback and conduction

G. M. Voit[1], M. Donahue[1], G. L. Bryan[2] & M. McDonald[3]

The Universe's largest galaxies reside at the centres of galaxy clusters and are embedded in hot gas that, if left undisturbed, would cool quickly and create many more new stars than are actually observed[1–5]. Cooling can be regulated by feedback from accretion of cooling gas onto the central black hole, but requires an accretion rate finely tuned to the thermodynamic state of the hot gas[6,7]. Theoretical models in which cold clouds precipitate out of the hot gas via thermal instability and accrete onto the black hole exhibit the necessary tuning[8–10]. Recent observational evidence shows that the abundance of cold gas in the centres of clusters increases rapidly near the predicted threshold for instability[11]. Here we report observations showing that this precipitation threshold extends over a large range in cluster radius, cluster mass and cosmic time. We incorporate the precipitation threshold into a framework of theoretical models for the thermodynamic state of hot gas in galaxy clusters. According to that framework, precipitation regulates star formation in some giant galaxies, while thermal conduction prevents star formation in others if it can compensate for radiative cooling and shut off precipitation.

Our framework can be expressed in terms of the time $t_{\rm cool}$ required for the hot gas to radiate an amount of energy equivalent to its current thermal energy. If intracluster gas were unable to cool, cosmological structure formation via hierarchical merging would produce galaxy clusters with radial cooling-time profiles that are similar to a baseline profile $t_{\rm base}(r)$, which can be computed with numerical simulations[12,13]. Massive galaxy clusters are observed to converge to this baseline profile at large radii[14], but radiative cooling cannot be ignored at smaller radii, where $t_{\rm cool}$ can be much shorter than the age of the Universe. Gas at small radii must either cool and condense or the cooling of that gas must trigger a thermal feedback that compensates for the radiative losses[15].

Thermal conduction is capable of compensating for cooling in cluster gas with $t_{\rm cool} > 1$ billion years (Gyr)[16,17]. Our framework therefore includes a locus of conductive balance, $t_{\rm cond}(r)$, along which thermal conduction exactly balances radiative cooling[18]. The locus itself is unstable, because conduction outcompetes cooling if $t_{\rm cool}$ is above that locus but cannot compete below it[19]. Conduction should therefore drive gas above the locus towards an isothermal core profile $t_{\rm iso}(r)$ identical to the baseline profile at large radii but with a constant temperature equal to the peak temperature of the baseline profile at smaller radii. Clusters in an isothermal core state have central cooling times exceeding $\sim$1 Gyr, and so mergers with other galaxy clusters, which occur on timescales of several billion years, can compete with cooling and further raise $t_{\rm cool}$ in the cores of those objects. Once $t_{\rm cool}$ exceeds the 14-Gyr age of the Universe, radiative cooling can no longer lower $t_{\rm cool}$, and this threshold corresponds to the 'no cooling' profile in our framework.

Clusters with cooling-time profiles that go below the locus of conductive balance require another heat source to balance cooling, and observations have shown that outflows emanating from a central supermassive black hole are sufficiently energetic to stop the cooling[7]. However, the triggering mechanism for that feedback response remained elusive until recent numerical simulations provided the missing puzzle piece[8–10,20,21]. Those simulations show that cold clouds start to precipitate out of hot-gas atmospheres in a state of global thermal balance when $t_{\rm cool}$ drops to ten times the free-fall time $t_{\rm ff} = (2r/g)^{1/2}$, where g is the local gravitational acceleration. The resulting precipitation feeds the central black hole through a 'chaotic cold accretion' process, producing a combination of thermal and kinetic feedback that maintains the necessary state of overall thermal balance[22,23]. Sporadic eruptions of feedback then cause the minimum value of $t_{\rm cool}/t_{\rm ff}$ to fluctuate within the range $5 < t_{\rm cool}/t_{\rm ff} < 20$.

We compute the critical profile for precipitation[24] by assuming a two-component gravitational potential. The first component is a mass-density profile $\propto (3r/r_{500})^{-1}[1+(3r/r_{500})]^{-2}$ in which the mean density within r_{500} is 500 times the cosmological critical density[25], and r_{500} depends on the cluster's gas temperature via $kT_{\rm X} \approx 125\mu m_{\rm p}[H(z)r_{500}]^2$, where $H(z)$ is the Hubble expansion parameter at the cluster's cosmological redshift z and $\mu m_{\rm p}$ is the mean mass per gas particle. The second component is a singular isothermal sphere (mass density $\propto r^{-2}$) with a velocity dispersion of 250 km s^{-1} to represent the stellar mass profile of the central galaxy[26]. Defining $t_{\rm precip}(r) = 10t_{\rm ff}$ then yields the critical profile for precipitation of cold clouds out of the hot gas.

There are thus three 'attractor' profiles for cluster cores: (1) dynamical heating via mergers will push hot gas towards a long-lived state with $t_{\rm cool} > 14$ Gyr, (2) thermal conduction will drive hot gas above the conductive-balance locus towards $t_{\rm iso}(r)$, and (3) hot gas below the conductive-balance locus will cool, sink into the central galaxy, fall into a precipitating state, and trigger feedback that prevents $t_{\rm cool}$ from dropping much below $10t_{\rm ff}$.

Comparing this framework of models with cooling-time profiles derived from the ACCEPT galaxy-cluster database[14] strongly supports the hypothesis that precipitation regulates cooling and star formation in massive galaxies (Fig. 1). The lower envelope of the $t_{\rm cool}(r)$ data closely follows the $\max[t_{\rm precip}(r), t_{\rm base}(r)]$ boundary over multiple orders of magnitude in radius, multiple orders of magnitude in cooling time, and more than an order of magnitude in system temperature. It even reproduces the kink at the intersection of $t_{\rm precip}(r)$ and $t_{\rm base}(r)$, confirming that the mechanism which regulates cooling and star formation in the Universe's largest galaxies prevents $t_{\rm cool}$ from dropping much below $10t_{\rm ff}$. This is an important finding, even if the precipitation-driven feedback model turns out to be incorrect, because it shows that the mechanism preventing runaway cooling in cluster cores depends critically on the $t_{\rm cool}/t_{\rm ff}$ ratio.

The data also imply that thermal conduction separates precipitating clusters from non-precipitating clusters because the locus of unstable conductive balance neatly divides systems with multiphase gas from those without it. Detections of Hα and far-infrared emission from cluster cores[14,27] indicate the presence of multiphase gas, and the cooling-time profiles of all multiphase cluster cores either drop below $t_{\rm cond}(r)$ or are in its vicinity. In contrast, nearly all of the clusters without observable Hα emission stay above $t_{\rm cond}(r)$. The few single-phase cluster cores that dip below $t_{\rm cond}(r)$ may be objects in transition to a precipitating state because they are still outside the precipitation zone at $5 < t_{\rm cool}/t_{\rm ff} < 20$. According to our framework, their multiphase counterparts with $20t_{\rm ff} < t_{\rm cool} < t_{\rm cond}(r)$ are likely to be systems in which a large burst of feedback has temporarily shut off precipitation but has not yet boosted

[1]Department of Physics and Astronomy, 567 Wilson Road, Michigan State University, East Lansing, Michigan 48864, USA. [2]Department of Astronomy, 1328 Pupin Physics Lab, MC 5246, 550 West 120th Street, New York 10027, USA. [3]MIT Kavli Institute for Astrophysics and Space Research, Massachusetts Institute of Technology (MIT), 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA.

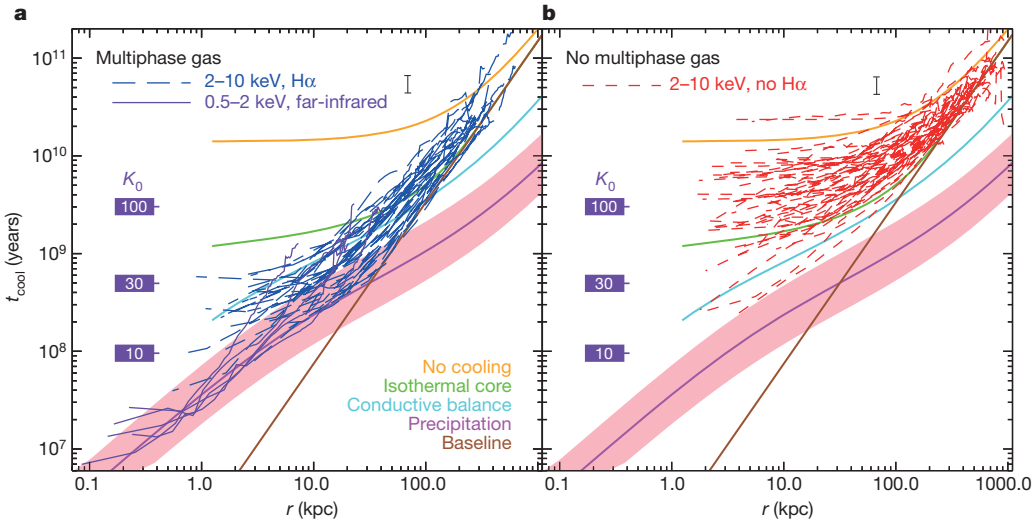


Figure 1 | Hot-gas cooling time as a function of radius in galaxy clusters. The observed ratio of cooling time to freefall time exhibits a hard floor at approximately 10, in accordance with model predictions[24] for precipitation-driven feedback. In **a**, dashed blue lines show cooling-time profiles for all objects in the ACCEPT database[14] with gas temperatures in the 2–10 keV range and Hα detections of multiphase gas. Solid purple lines show all 0.5–2.0 keV objects in ACCEPT with far-infrared detections of multiphase gas. The lower envelope of the cooling-time profiles closely follows the boundary defined by the precipitation threshold at $t_{cool}/t_{ff} \approx 10$ (thick magenta line) and the cosmological baseline profile (brown), and most of those profiles enter the zone at $5 < t_{cool}/t_{ff} < 20$ (pink), within which precipitation-driven feedback stabilizes simulated galaxy clusters. The upper end of the t_{cool} envelope for multiphase systems lies in the vicinity of the locus of unstable conductive balance (cyan), indicating that thermal conduction eliminates multiphase gas above that locus. In **b**, dashed red lines show cooling-time profiles for all 2–10 keV objects in the ACCEPT database and no observable Hα emission. None of those profiles enters the precipitation zone, nearly all are above the locus of conductive balance, and most are between the isothermal core profile (green) and the cooling threshold (orange) at which the minimum cooling time equals the age of the Universe. All of the thick solid lines show model predictions for a 6 keV cluster, and purple tags indicate the core entropy index (K_0 in keV cm^2) at this temperature. An error bar near the upper right corner shows the typical uncertainty range (2 s.d.) for t_{cool}, which comes primarily from the statistical uncertainty in gas temperatures derived from Chandra X-ray spectroscopy.

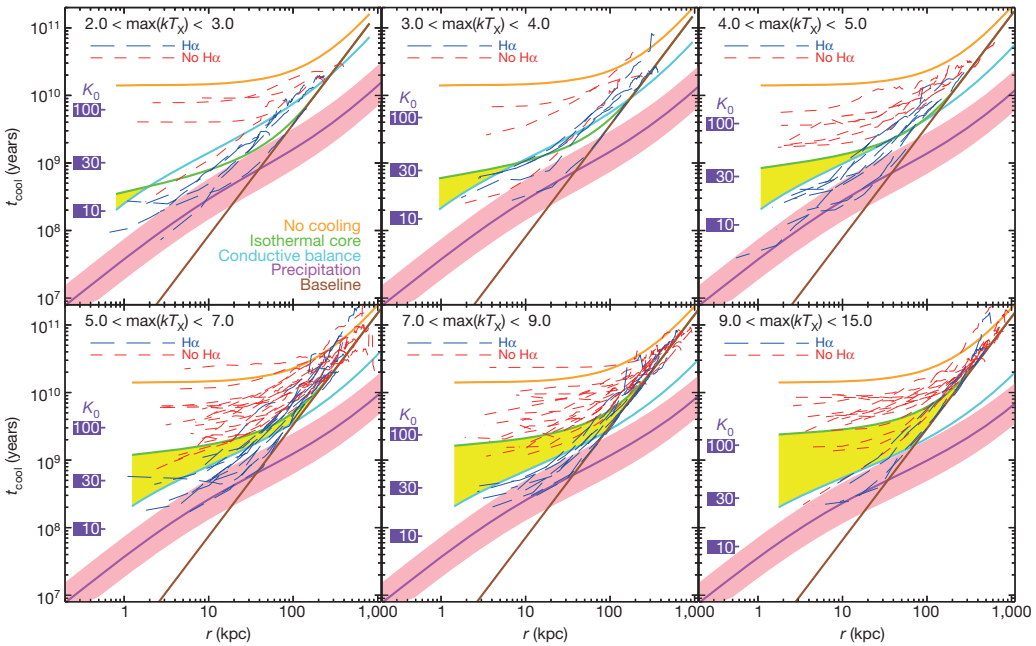


Figure 2 | Hot-gas cooling time as a function of radius in galaxy clusters of differing temperatures. All lines are colour-coded as in Fig. 1. When grouped by temperature, all of the Hα-emitting clusters have profiles that dip below the locus of conductive balance, while only three of the no-Hα clusters dip below it. None of those three enters the pink zone corresponding to the t_{cool}/t_{ff} excursions seen in simulations of precipitation-driven feedback, suggesting that the three clusters may be objects in which precipitation has not yet begun. In the yellow regions, our model predicts that thermal conduction should be heating gas and driving it to the isothermal-core state. If thermal conduction is indeed responsible for separating the t_{cool} profiles of Hα and no-Hα clusters, then the degree of separation should increase with temperature. The main effect of increasing temperature is to drive the locus of conductive balance closer to the precipitation threshold, narrowing the range of t_{cool}/t_{ff} within which multiphase gas can persist. This trend appears to be present in the data but with marginal statistical significance. For Hα-emitting clusters in the 2–7 keV range, we find that the mean value of $\min[t_{cool}/t_{ff}]$ is 20.9 ± 1.7 with a standard deviation of 9.5. Among Hα-emitting clusters in the 7–15 keV range, both the average value of $\min[t_{cool}/t_{ff}]$ and the dispersion are lower, with a mean of 15.7 ± 1.7 and a standard deviation of 5.6.

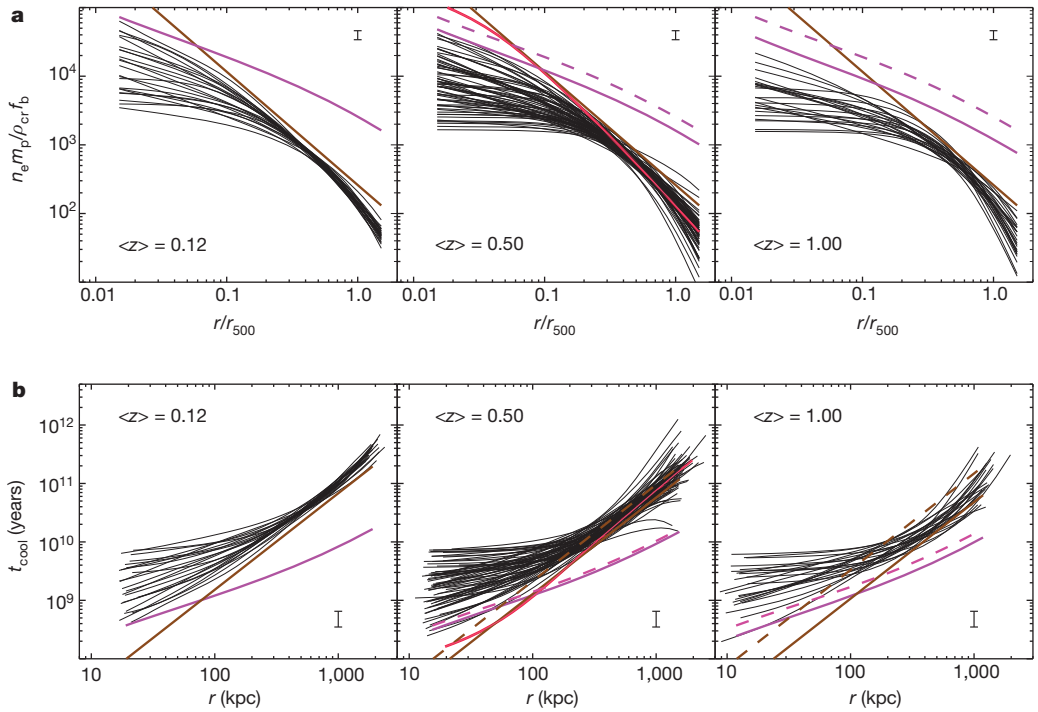


Figure 3 | Evolution of radial gas density and cooling-time profiles in galaxy clusters. **a**, Evolution of electron density n_e. Objects at mean cosmological redshift $<z> = 0.12$ are from the Chandra Cluster Cosmology Project[29], and objects at $<z> = 0.5$ and 1.0 are from the South Pole Telescope[28] survey. Cosmological scaling has been removed through division of r by r_{500} and division of n_e by $\rho_{cr} f_b m_p^{-1}$, where ρ_{cr} is the cosmological critical density and f_b is the fraction of cosmic mass in baryonic form. Thin lines show cluster observations. Solid thick magenta and brown lines show the precipitation limit and baseline profile, respectively, corresponding to a reference temperature of 6 keV. Dashed lines show the precipitation and baseline profiles for the low-redshift subsample at $<z> = 0.12$. The gas-density contrast between a core near the precipitation limit and the outer part of the baseline profile decreases with increasing redshift. This happens because the Universe as a whole is denser at earlier times, whereas gas density at the precipitation limit remains nearly constant because it is set by local conditions. **b**, Evolution of hot-gas cooling time. All line styles are identical to those in **a**. In this unscaled representation of the same data, the precipitation limit remains nearly constant, while the baseline profile shifts downward with increasing redshift because the mean gas density is increasing. Error bars in both panels show a statistical uncertainty range equivalent to 2 s.d. One South Pole Telescope cluster in the $<z> = 0.5$ set, shown with a red line, crosses the precipitation limit. Notably, it is the Phoenix cluster[30], which has, by far, the largest central star-formation rate of all known galaxy clusters.

t_{cool} high enough for conduction to eliminate the multiphase gas. Those cluster cores should cool and return to active precipitation within a few hundred million years.

Subdividing clusters according to temperature strengthens the case for thermal conduction (Fig. 2). Our framework predicts that clusters with cooling-time profiles between the conductive-balance locus and the isothermal-core profile should be rare, because thermal conduction should be driving t_{cool} from $t_{cond}(r)$ towards $t_{iso}(r)$. The data show that the zone between $t_{cond}(r)$ and $t_{iso}(r)$ is indeed systematically depopulated and suggest that it grows larger with increasing temperature, in accord with the strong temperature dependence of thermal conduction in astrophysical plasmas. We note also that in every temperature range, the lower edge of the t_{cool} envelope closely follows the joint precipitation + baseline profile, including the kink at the intersection point, showing that the floor at $t_{cool} \approx 10 t_{ff}$ is present in data across the entire cluster temperature range.

Two predictions for the evolution of galaxy-cluster cores follow from these considerations. First, the thermodynamic properties of precipitating cores should remain relatively constant with time, because they are determined by local conditions and not by cosmological evolution. Second, the contrast in gas density between a precipitating core and the outer parts of a cluster should grow more pronounced with time, because hierarchical structure formation causes the baseline profile to become less dense as dynamical heating resulting from mergers adds entropy to the gas and shifts t_{cool} upward. X-ray observations of the South Pole Telescope galaxy-cluster sample[28] support these predictions (Fig. 3). The limits on central density, entropy, and cooling time of high-redshift clusters remain similar to those for low-redshift clusters and do not violate the precipitation limit, whereas the outer parts remain limited by the baseline profile, which is at progressively greater density, lower entropy, and shorter cooling time as cluster redshift increases.

Taken as a whole, this many-faceted correspondence between models and data convincingly shows that we now understand what regulates cooling and star formation in the Universe's largest galaxies and raises an even bigger question. How far down the galaxy-mass spectrum do these principles extend? Precipitation is likely to be a very general feature of galaxy evolution, in that precipitation-driven feedback owing to both star formation and accretion onto black holes is likely to maintain the ambient circumgalactic medium of a star-forming galaxy in a state with $t_{cool}/t_{ff} \approx 10$. Conversely, galaxies embedded in ambient gas with $t_{cool}/t_{ff} \gg 10$ have no way of replenishing the cold gas required for star formation, which therefore wanes. Thermal conduction is probably less general, given its strong temperature dependence, but stellar heating mechanisms such as supernova explosions should be of greater relative importance in lower-temperature systems and may provide an analogous upper bound on residual precipitation that separates star-forming galaxies from those in which star formation has ceased.

Received 3 September; accepted 22 December 2014.
Published online 4 March 2015.

1. Fabian, A. C. Cooling flows in clusters of galaxies. *Annu. Rev. Astron. Astrophys.* **32,** 277–318 (1994).
2. Borgani, S. *et al.* X-ray properties of galaxy clusters and groups from a cosmological hydrodynamics simulation. *Mon. Not. R. Astron. Soc.* **348,** 1078–1096 (2004).
3. Peterson, J. R. & Fabian, A. C. X-ray spectroscopy of cooling clusters. *Phys. Rep.* **427,** 1–39 (2006).

4. Nagai, D., Kravtsov, A. V. & Vikhlinin, A. Effects of galaxy formation on the thermodynamics of the intracluster medium. *Astrophys. J.* **668,** 1–14 (2007).
5. O'Dea, C. P. *et al.* An infrared survey of brightest cluster galaxies. II. Why are some brightest cluster galaxies forming stars? *Astrophys. J.* **681,** 1035–1045 (2008).
6. Soker, N., White, R. E., David, L. P. & McNamara, B. R. A moderate cluster cooling flow model. *Astrophys. J.* **549,** 832–839 (2001).
7. McNamara, B. R. & Nulsen, P. E. J. Mechanical feedback from active galactic nuclei in galaxies, groups, and clusters. *New J. Phys.* **14,** 055023 (2012).
8. McCourt, M., Sharma, P., Quataert, E. & Parrish, I. Thermal instability in gravitationally-stratified plasma: implications for multiphase structure in clusters and galaxy haloes. *Mon. Not. R. Astron. Soc.* **419,** 3319–3337 (2012).
9. Sharma, P., McCourt, M., Quataert, E. & Parrish, I. Thermal instability and the feedback regulation of hot halos in clusters and groups of galaxies. *Mon. Not. R. Astron. Soc.* **420,** 3174–3194 (2012).
10. Gaspari, M., Ruszkowski, M. & Sharma, P. Cause and effect of feedback: multiphase gas in cluster cores heated by AGN jets. *Astrophys. J.* **746,** 94–108 (2012).
11. Voit, G. M. & Donahue, M. Cooling time, freefall time, and precipitation in the cores of ACCEPT galaxy clusters. *Astrophys. J.* **799,** L1 (2015).
12. Frenk, C. S. *et al.* The Santa Barbara Cluster Comparison Project. *Astrophys. J.* **525,** 554–582 (1999).
13. Voit, G. M., Kay, S. T. & Bryan, G. L. The baseline intracluster entropy profile from gravitational structure formation. *Mon. Not. R. Astron. Soc.* **364,** 909–916 (2005).
14. Cavagnolo, K. W., Donahue, M., Voit, G. M. & Sun, M. Intracluster medium entropy profiles for a Chandra archival sample of galaxy clusters. *Astrophys. J. Suppl. Ser.* **182,** 12–32 (2009).
15. Voit, G. M. & Bryan, G. L. Regulation of the X-ray luminosity of clusters of galaxies by cooling and supernova feedback. *Nature* **414,** 425–427 (2001).
16. Zakamska, N. L. & Narayan, R. Models of galaxy clusters with thermal conduction. *Astrophys. J.* **582,** 162–169 (2003).
17. Voigt, L. M. & Fabian, A. C. Thermal conduction and reduced cooling flows in galaxy clusters. *Mon. Not. R. Astron. Soc.* **347,** 1130–1149 (2004).
18. Voit, G. M. Quasi-steady configurations of conductive intracluster media. *Astrophys. J.* **740,** 28–38 (2011).
19. Bregman, J. N. & David, L. P. Heat conduction in cooling flows. *Astrophys. J.* **326,** 639–644 (1988).
20. Li, Y. & Bryan, G. L. Modeling active galactic nucleus feedback in cool-core clusters: the balance between heating and cooling. *Astrophys. J.* **789,** 54–67 (2014).
21. Li, Y. & Bryan, G. L. Modeling active galactic nucleus feedback in cool-core clusters: the formation of cold clumps. *Astrophys. J.* **789,** 153–164 (2014).
22. Gaspari, M., Ruszkowski, M. & Oh, S. P. Chaotic cold accretion onto black holes. *Mon. Not. R. Astron. Soc.* **432,** 3401–3422 (2013).
23. Gaspari, M., Ruszkowski, M., Oh, S. P., Brighenti, F. & Temi, P. Chaotic cold accretion onto black holes in rotating atmospheres. Preprint at http://arxiv.org/abs/1407.7531 (2014).
24. Sharma, P., McCourt, M., Parrish, I. & Quataert, E. On the structure of hot gas in halos: implications for the L_X-T_X relation and missing baryons. *Mon. Not. R. Astron. Soc.* **427,** 1219–1228 (2012).
25. Navarro, J. F., Frenk, C. S. & White, S. D. M. A universal density profile from hierarchical clustering. *Astrophys. J.* **490,** 493–508 (1997).
26. Bernardi, M. *et al.* The luminosities, sizes, and velocity dispersions of brightest cluster galaxies: implications for formation history. *Astron. J.* **133,** 1741–1755 (2007).
27. Hoffer, A. S., Donahue, M., Hicks, A. & Barthelemy, R. S. Infrared and ultraviolet star formation in brightest cluster galaxies in the ACCEPT sample. *Astrophys. J. Suppl. Ser.* **199,** 23–38 (2012).
28. McDonald, M. *et al.* The growth of cool cores and evolution of cooling properties in a sample of 83 galaxy clusters at $0.3<z<1.2$ selected from the SPT-SZ survey. *Astrophys. J.* **774,** 23–45 (2013).
29. Vikhlinin, A. *et al.* Chandra Cluster Cosmology Project. II. Samples and X-ray data reduction. *Astrophys. J.* **692,** 1033–1059 (2009).
30. McDonald, M. *et al.* A massive, cooling-flow-induced starburst in the core of a luminous cluster of galaxies. *Nature* **774,** 23–45 (2012).

Acknowledgements G.M.V. and M.D. acknowledge NSF support through grant AST-0908819. G.L.B. acknowledges NSF AST-1008134, AST-1210890, NASA grant NNX12AH41G, and XSEDE Computational resources. M.McD. acknowledges support by NASA through a Hubble Fellowship grant HST-HF51308.01-A awarded by the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Inc., for NASA, under contract NAS 5-26555.

Author Contributions G.M.V.: theoretical models, data interpretation, manuscript preparation; M.D.: data analysis, data interpretation, discussions, manuscript review; G.L.B.: theoretical models, discussions, manuscript review; M.McD.: data analysis, discussions, manuscript preparation, manuscript review.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to G.M.V. (voit@pa.msu.edu).

LETTER

doi:10.1038/nature14262

Ongoing hydrothermal activities within Enceladus

Hsiang-Wen Hsu[1]*, Frank Postberg[2,3]*, Yasuhito Sekine[4]*, Takazo Shibuya[5], Sascha Kempf[1], Mihály Horányi[1], Antal Juhász[1,6], Nicolas Altobelli[7], Katsuhiko Suzuki[8], Yuka Masaki[8], Tatsu Kuwatani[9], Shogo Tachibana[10], Sin-iti Sirono[11], Georg Moragas-Klostermeyer[3] & Ralf Srama[3]

Detection of sodium-salt-rich ice grains emitted from the plume of the Saturnian moon Enceladus suggests that the grains formed as frozen droplets from a liquid water reservoir that is, or has been, in contact with rock[1,2]. Gravitational field measurements suggest a regional south polar subsurface ocean of about 10 kilometres thickness located beneath an ice crust 30 to 40 kilometres thick[3]. These findings imply rock–water interactions in regions surrounding the core of Enceladus. The resulting chemical 'footprints' are expected to be preserved in the liquid and subsequently transported upwards to the near-surface plume sources, where they eventually would be ejected and could be measured by a spacecraft[4]. Here we report an analysis of silicon-rich, nanometre-sized dust particles[5–8] (so-called stream particles) that stand out from the water-ice-dominated objects characteristic of Saturn. We interpret these grains as nanometre-sized SiO_2 (silica) particles, initially embedded in icy grains emitted from Enceladus' subsurface waters and released by sputter erosion in Saturn's E ring. The composition and the limited size range (2 to 8 nanometres in radius) of stream particles indicate ongoing high-temperature (>90 °C) hydrothermal reactions associated with global-scale geothermal activity that quickly transports hydrothermal products from the ocean floor at a depth of at least 40 kilometres up to the plume of Enceladus.

Dust dynamics provide diagnostic information about the origin of the observed dust populations. The dynamical properties of Saturnian stream particles show characteristics inherited from Saturn's diffuse E ring[7]. Considering the long-term evolution of the E ring and dust–plasma interactions, our dynamical analysis reproduces the observed characteristics, confirming their E-ring origin (Methods). Enceladus is the source of the E ring and hence the ultimate source of stream particles, allowing Enceladus to be probed using stream particle measurements.

Co-added mass spectra of selected Saturnian stream particles detected by Cassini's Cosmic Dust Analyser (CDA)[9] (Fig. 1) show silicon as the only highly significant particle constituent. Oxygen is the other abundant possible particle mass line but is also a minor but frequent target contaminant[10]. The contribution of particle material to the oxygen signal is difficult to assess, but its intensity is in agreement with at least a fractional contribution from silicates (Methods). Remarkably, only traces (at most) of metals are found to contribute to the particle composition, indicating that the stream particle spectra are not in agreement with those of typical rock-forming silicate minerals (that is, olivine or pyroxene). The data are in agreement solely with extremely metal-poor (or metal-free) silicon-bearing compounds, of which, besides elemental Si, only SiO_2 and SiC are of cosmochemical relevance[11]. Considering that Si and SiC are highly unlikely to be emitted in significant quantities from a planetary body, we conclude that the dominant, if not sole, constituent of most stream particles must therefore be SiO_2. Quantitative mass spectra analysis indicates a radius of $r_{max} = 6–9$ nm for the largest stream particles (Methods). This is in excellent agreement with the upper particle size limit independently inferred from dynamical simulations ($r_{max} \approx 8$ nm)[7].

The spontaneous, homogeneous nucleation of nanometre-sized colloidal silica is a unique property of the silica–water system. We consider this as the production mechanism of the observed silica nanoparticles because of (1) the existence of a subsurface ocean in contact with rock and (2) the improbability of homogeneous fragmentation of pure bulk silica into particles with radii exclusively below 10 nm within Enceladus. Only a rock-related, bottom-up formation process is plausible. Colloidal silica nanoparticles form with initial radii of 1–1.5 nm when the solution becomes supersaturated[12]. In moderately alkaline solutions (pH 7.5–10.5) with low electrolyte concentration, the charge state of silica nuclei allows colloidal silica nanoparticles to nucleate and grow by addition of dissolved silica as well as by Ostwald ripening[12,13]. Above about pH 10.5, silica solubility becomes too high to maintain a stable colloidal phase[12]. Laboratory experiments show that after hours to days in a supersaturated solution with a slightly alkaline pH and at various ionic strengths, colloidal silica grows to radii of 2–6 nm (refs 14–17), which is in good agreement with CDA measurements.

Both measurements—mass spectra and the narrow size distribution—indicate silica nanoparticles but may not provide unequivocal proof

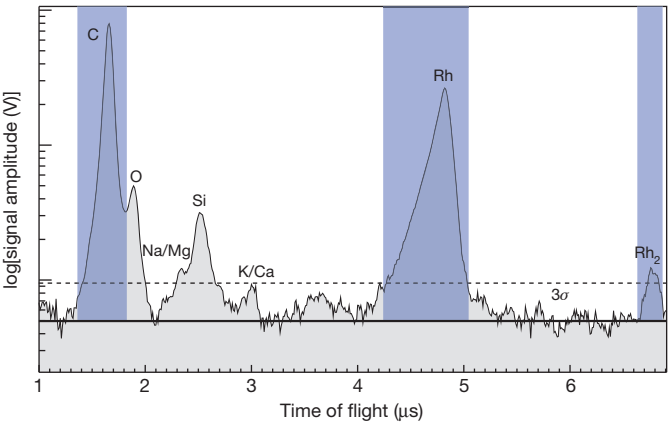


Figure 1 | Identifying particle constituents. Shown is a co-added impact ionization mass spectrum from 32 selected Saturnian stream particle spectra with the strongest Si^+ signals. As expected, the impacts produce more ions from the CDA's target material (Rh^+ and Rh_2^+; blue areas) and the target contaminants[6,10] (C^+, H^+; blue areas, H^+ not shown) than from the nanoparticle itself. Ions O^+ and Si^+ are the most abundant potential particle mass lines. Na^+/Mg^+ (solidus indicates the two species can not be distinguished) form the only other potential particle mass line with a signal-to-noise ratio above 3σ (dashed line; σ, standard deviation). The particle composition agrees best with pure silica when the target impurities and the impact ionization process are taken into account (Methods).

[1]Laboratory for Atmospheric and Space Physics, University of Colorado, Boulder, Colorado 80303, USA. [2]Institut für Geowissenschaften, Universität Heidelberg, 69120 Heidelberg, Germany. [3]Institut für Raumfahrtsysteme, Universität Stuttgart, 70569 Stuttgart, Germany. [4]Department of Complexity Science and Engineering, University of Tokyo, Kashiwa 277-8561, Japan. [5]Laboratory of Ocean–Earth Life Evolution Research, JAMSTEC, Yokosuka 237-0061, Japan. [6]Institute for Particle and Nuclear Physics, Wigner RCP, 1121 Budapest, Hungary. [7]European Space Agency, ESAC, E-28691 Madrid, Spain. [8]Research and Development Center for Submarine Resources, JAMSTEC, Yokosuka 237-0061, Japan. [9]Graduate School of Environmental Studies, Tohoku University, Sendai 980-8579, Japan. [10]Department of Natural History Sciences, Hokkaido University, Sapporo 060-0810, Japan. [11]Graduate School of Environmental Sciences, Nagoya University, Nagoya 464-8601, Japan.
*These authors contributed equally to this work.

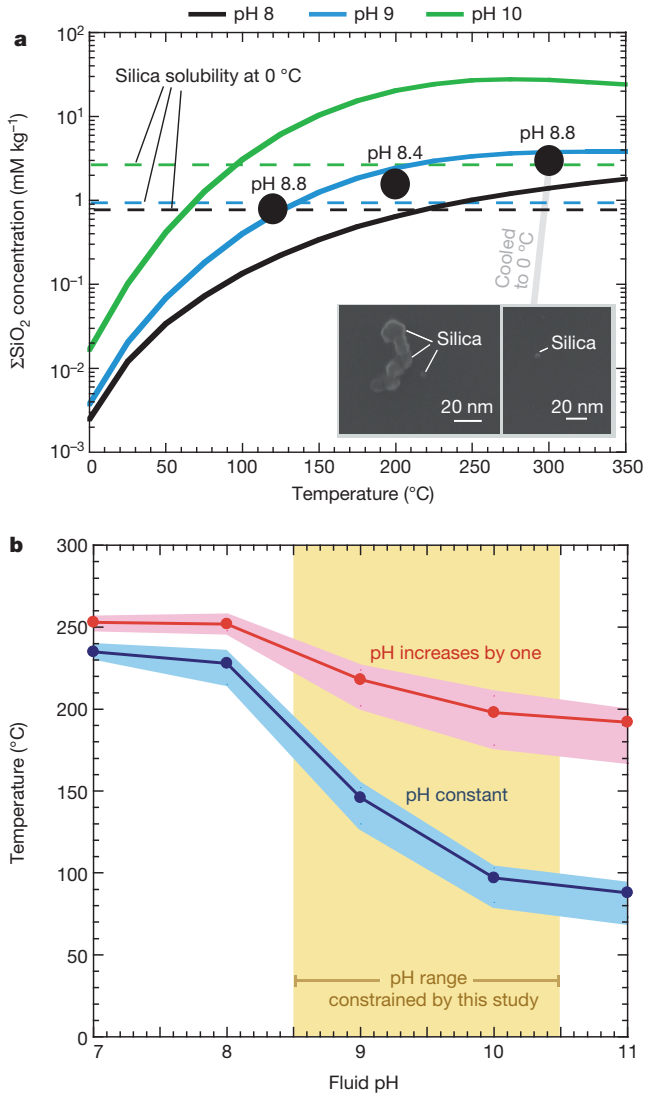


Figure 2 | Minimum temperatures for formation of silica nanoparticles. **a**, Solid lines show ΣSiO_2 of a serpentine–talc/saponite buffer equilibrium as a function of temperature (*x* axis) and pH (line colour: see key above). This buffer system is consistent with the measured ΣSiO_2 in fluid samples of the hydrothermal experiments using an orthopyroxene and olivine powder mixture at 400-bar pressure (filled black circles annotated with *in situ* pH values; Methods). Dashed lines show the 0 °C silica solubility at the respective pH. The difference between the solid and dashed lines determines the amount of ΣSiO_2 available for silica nanoparticle formation at the respective pH. Insets, images of silica nanoparticles formed in cooled solutions. **b**, Relationships between minimum hydrothermal fluid temperatures and fluid pH for silica nanoparticle formation. Red and blue colours represent results with increasing and fixed pH, respectively, upon cooling and mixing with seawater. Data points show results for Na^+ concentration 0.1 mol kg^{-1} and pressure 30 bar; shaded areas represent the uncertainties in Na^+ concentrations (0.05–0.3 mol kg^{-1}) and pressure (10–80 bar; ref. 3).

individually. However, nanosized silica remains as the only plausible interpretation of the stream particle measurements when results from these two independent analysis methods are incorporated. Moreover, the relation between the stability of silica nanoparticles and solvent alkalinity matches the pH range of the liquid plume source(s) (about 8.5–9), as independently inferred from the composition of emitted salt-rich ice grains[1,2].

We can now use silica nanoparticles as a thermometer for the subsurface ocean floor of Enceladus, assuming that such particles form owing to SiO_2 solubility reduction during a temperature reduction in cooling water[17,18]. This is the most common way for silica nanoparticles to form on Earth, and is frequently observed in alkaline hydrothermal fluids[12,17–19]. To determine the relation of silica concentration versus solution temperature applicable to Enceladus, long-term rock–water interaction experiments were conducted. A pressurized solution of $NaHCO_3$ and NH_3 in water was brought into contact with powdered primordial minerals (70% olivine and 30% pyroxene) at various temperatures and for several months (Methods). Hydrothermal alterations produced secondary minerals typically found in carbonaceous chondrites, including serpentine, talc/saponite and magnetite. Our experimental results (Fig. 2a) show that the total SiO_2 concentration in fluids ($\Sigma SiO_2 = SiO_2(aq.) + HSiO_3^- + NaSiO_3(aq.)$) in contact with these secondary minerals is controlled by a serpentine–talc/saponite buffer system: that is, serpentine + $2SiO_2(aq.) \leftrightarrow$ talc/saponite + H_2O. This allows us to calculate the minimum temperature required for silica nanoparticle formation on cooling of the hydrothermal fluids—that is, the reaction temperature at which ΣSiO_2 exceeds the solubility of amorphous silica at 0 °C for a given pH. Assuming the fluid pH remains constant on cooling, the reaction temperature must reach ~90 °C at pH 10.5, or a higher temperature if the fluid pH is below 10.5 (Fig. 2b). Because silica solubility increases with fluid alkalinity, the minimum temperature allowing silica nanoparticle formation on subsequent cooling rises to ~190 °C at pH 10.5 if the hydrothermal fluid pH were to increase by one when mixing with the subsurface ocean water (Methods and ref. 20).

It is not clear how steep the temperature gradient across the subsurface ocean is. However, the ocean is most likely to be convective if the minimum temperature allowing silica nanoparticle formation on subsequent cooling (that is, > 90 °C) at the rock–water interface is achieved. We believe that most silica nucleation and initial growth would occur when the hydrothermal fluids reach the relatively cold ocean water at the ocean floor. The growth of silica nanoparticles may continue as the hydrothermal fluids ascend (Fig. 3).

For comparison, the average concentration of silica nanoparticles in their icy E-ring 'carrier grains' can be estimated using the measured and modelled stream particle production rate (Fig. 4 and Methods). Albeit with large uncertainties, a conservative lower limit still requires the formation of 150 p.p.m. of silica nanoparticles, equivalent to a solution supersaturated by about 2.5 mM SiO_2, which was available to form the observed nanoparticles. Such a high nanosilica abundance requires a high temperature gradient at a pH of at least 8.5, and cannot be explained solely by incorporation of dissolved silica on freezing of water droplets in the vents[2]. The high abundance and specific sizes of stream particles both indicate that they existed in colloidal form before their integration into ice grains.

The existence of silica nanoparticles also provides strict constraints on the salinity of Enceladus' subsurface waters because silica colloids aggregate and precipitate quickly at high ionic strength[12,13]. The critical coagulation concentration of NaCl at pH 9 is 2% or ~0.3 M (1.5% or ~0.2 M at pH 10, 4% or ~0.6 M at pH 8)[13]. This sets an upper salinity limit of about 4% for the location where silica nanoparticles form at depth, as well as for the near-surface plume sources, and corresponds to the lower salinity limit of 0.5% derived earlier[1]. Partial freezing of the water would increase the salinity and would result in immediate silica precipitation[19], suggesting that the observed silica nanoparticles have never 'seen' a brine. This also implies that the observed silica nanoparticles were produced during the present active phase of Enceladus.

The growth of colloidal particles sets another constraint on the lifetime of the silica nanoparticles. For example, through Ostwald ripening[21], nanosilica would grow to micrometre-sized grains within a few thousand years or less (Methods). The observed radii, below 10 nm, imply the continuous and relatively fast upward transportation of hydrothermal products (see, for example, ref. 22), from ongoing hydrothermal activities in the subsurface ocean to the plume sources close to the surface, over months to several years at most (Methods).

Our results show that two very different dust populations detected by Cassini—that is, micrometre-sized ice grains[1,2,4,23,24] and nanometre-sized

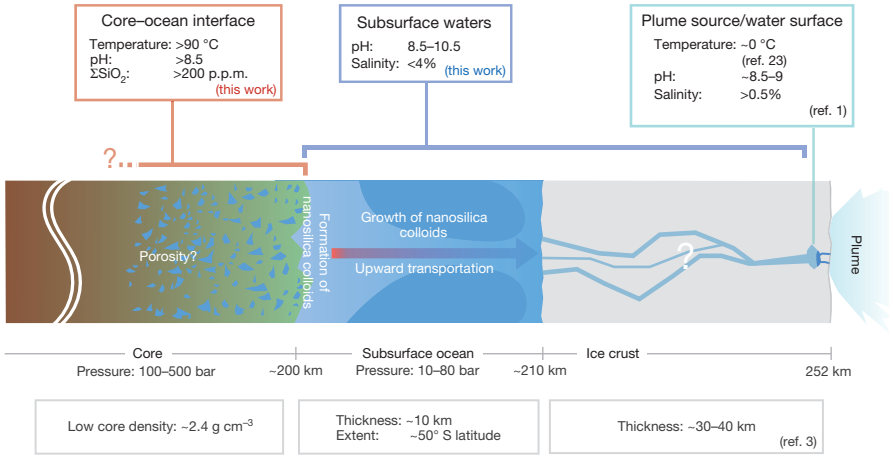


Figure 3 | A schematic of Enceladus' interior. The internal structure and conditions of Enceladus beneath its south polar region derived from this and previous work. The main components (core, subsurface ocean, ice crust and plume) are shown left to right; top row gives temperature and chemical properties of each component, middle row shows schematic structure, and bottom row gives physical properties. Distances labelling the grey line below the middle row are distances from the centre of Enceladus towards its south pole (not to scale).

silica stream particles—in fact have the same origin but probe the conditions of the subsurface water of Enceladus at different depths: the silica nanoparticles probe the pH, salinity and water temperature at the bottom of Enceladus' ocean, while the micrometre-sized ice dust grains reveal composition and thermal dynamical processes at near-surface liquid plume sources and in the vents[1,2,23] (Fig. 3). The current plume activity is probably not superficial but a large, core-to-surface-scale process. The low core densities implied by Cassini's gravitational field measurements[3] as well as the low pressure of the mantle resting on the core[25] are in good agreement with a porous core. This would allow water to percolate through it, providing a huge surface area for rock–water interactions, and the high temperatures (>90 °C) implied by our observations might occur deep inside Enceladus' core.

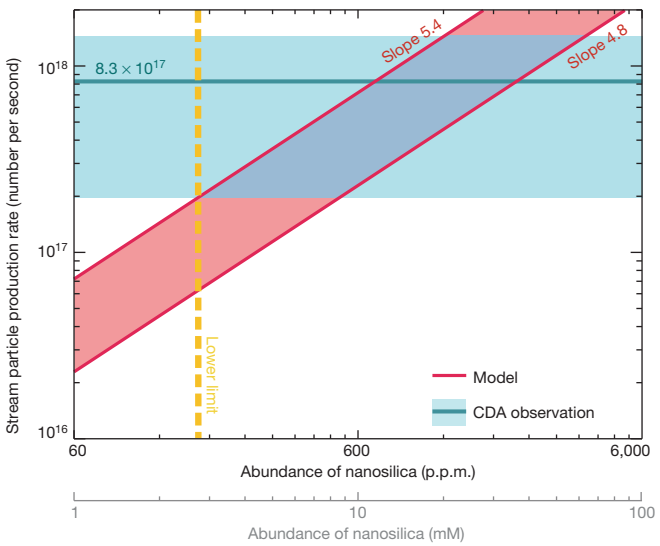


Figure 4 | Concentration of silica nanoparticles in E-ring grains. The mass fraction of silica nanoparticles in E-ring ice grains is estimated by comparing the production rates derived from the dynamical model (sloping red lines) and CDA measurements (blue horizontal line and shaded region). We assume that the stream particle release rate is directly proportional to the E-ring sputtering erosion rate. The steeper the power-law size distribution slope (μ), the larger the total surface area of E-ring grains and thus the higher the production rate of silica nanoparticles. The lower limit for the nanosilica mass fraction is ~150 p.p.m. (equivalent to 2.5 mM shown in the lower x axis) with $\mu = 5.4$ (yellow dashed line)[24].

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 29 November 2013; accepted 26 January 2015.

1. Postberg, F. *et al.* Sodium salts in E-ring ice grains from an ocean below the surface of Enceladus. *Nature* **459,** 1098–1101 (2009).
2. Postberg, F., Schmidt, J., Hillier, J., Kempf, S. & Srama, R. A salt-water reservoir as the source of a compositionally stratified plume on Enceladus. *Nature* **474,** 620–622 (2011).
3. Iess, L. *et al.* The gravity field and interior structure of Enceladus. *Science* **344,** 78–80 (2014).
4. Hillier, J. K. *et al.* The composition of Saturn's E ring. *Mon. Not. R. Astron. Soc.* **377,** 1588–1596 (2007).
5. Kempf, S. *et al.* High-velocity streams of dust originating from Saturn. *Nature* **433,** 289–291 (2005).
6. Kempf, S. *et al.* Composition of Saturnian stream particles. *Science* **307,** 1274–1276 (2005).
7. Hsu, H.-W. *et al.* Stream particles as the probe of the dust-plasma-magnetosphere interaction at Saturn. *J. Geophys. Res.* **116,** A09215 (2011).
8. Hsu, H.-W., Krüger, H. & Postberg, F. in *Nanodust in the Solar System: Discoveries and Interpretations* (eds Mann, I., Meyer-Vernet, N. & Czechowski, A.) 77–117 (Springer Astrophysics and Space Science Library, Vol. 385, 2012).
9. Srama, R. *et al.* The Cassini cosmic dust analyzer. *Space Sci. Rev.* **114,** 465–518 (2004).
10. Postberg, F. *et al.* Discriminating contamination from particle components in spectra of Cassini's dust detector CDA. *Planet. Space Sci.* **57,** 1359–1374 (2009).
11. Ming, T. *et al.* Meteoritic silicon carbide and its stellar sources — implications for galactic chemical evolution. *Nature* **339,** 351–354 (1989).
12. Iler, R. K. *The Chemistry of Silica* (Wiley & Sons, 1979).
13. Allen, L. H. & Matijević, E. Stability of colloidal silica, I. Effect of simple electrolytes. *J. Colloid Interface Sci.* **31,** 287–296 (1969).
14. Icopini, G. A., Brantley, S. L. & Heaney, P. J. Kinetics of silica oligomerization and nanocolloid formation as a function of pH and ionic strength at 25°C. *Geochim. Cosmochim. Acta* **69,** 293–303 (2005).
15. Conrad, C. F. *et al.* Modeling the kinetics of silica nanocolloid formation and precipitation in geologically relevant aqueous solutions. *Geochim. Cosmochim. Acta* **71,** 531–542 (2007).
16. Tobler, D. J., Shaw, S. & Benning, L. G. Quantification of initial steps of nucleation and growth of silica nanoparticles: an *in-situ* SAXS and DLS study. *Geochim. Cosmochim. Acta* **73,** 5377–5393 (2009).
17. Tobler, D. J. & Benning, L. G. *In situ* and time resolved nucleation and growth of silica nanoparticles forming under simulated geothermal conditions. *Geochim. Cosmochim. Acta* **114,** 156–168 (2013).
18. Herzig, P. M. *et al.* Hydrothermal silica chimney fields in the Galapagos Spreading Center at 86° W. *Earth Planet. Sci. Lett.* **89,** 261–272 (1988).
19. Channing, A. & Butler, I. B. Cryogenic opal-A deposition from Yellowstone hot springs. *Earth Planet. Sci. Lett.* **257,** 121–131 (2007).
20. Zolotov, M. Y. Aqueous fluid composition in CI chondritic materials: chemical equilibrium assessments in closed systems. *Icarus* **220,** 713–729 (2012).
21. Sirono, S. Differentiation of silicates from H_2O ice in an icy body induced by ripening. *Earth Planets Space* **65,** 1563–1568 (2013).
22. Matson, D. L., Castillo-Rogez, J. C., Davies, A. G. & Johnson, T. V. Enceladus: a hypothesis for bringing both heat and chemicals to the surface. *Icarus* **221,** 53–62 (2012).
23. Schmidt, J., Brilliantov, N., Spahn, F. & Kempf, S. Slow dust in Enceladus' plume from condensation and wall collisions in tiger stripe fractures. *Nature* **451,** 685–688 (2008).

24. Kempf, S. *et al.* The E ring in the vicinity of Enceladus. I. Spatial distribution and properties of the ring particles. *Icarus* **193,** 420–437 (2008).
25. Malamud, U. & Prialnik, D. Modeling serpentinization: applied to the early evolution of Enceladus and Mimas. *Icarus* **225,** 763–774 (2013).

Acknowledgements We acknowledge support from the CDA team, the Cassini project, and NASA. This research was partly supported by Grants-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology, Japan, by the Japan Society for the Promotion of Science, and by the Astrobiology Program of the National Institutes of Natural Sciences, Japan. This work was partly supported by the DLR grant 50 OH1103. We thank J. Schmidt, M. Y. Zolotov and D. J. Tobler for discussions, and E. S. Guralnick, J. K. Hillier, A. Rasca and T. Munsat for advice on writing this Letter. Y.S. thanks A. Okubo for her technical help in taking FE-SEM images.

Author Contributions H.-W.H., F.P. and Y.S. outlined the study and wrote the Letter. H.-W.H. performed the CDA dynamic analyses with assistance from A.J., M.H. and S.K.; F.P. and S.K. performed the CDA composition analyses; S.K., G.M.-K. and R.S. performed the CDA measurements and initial data processing; F.P. and N.A. performed the CDA mass spectra data acquisition and data reduction; Y.S. performed the experiments and calculations simulating Enceladus' ocean conditions; T.S. designed the hydrothermal experiments and the analysis system; S.T. synthesized starting minerals for the experiments; K.S., Y.M. and T.K. contributed to performing fluid and solid analyses in the experiments; and S.-i.S. estimated the lifetime of silica nanoparticles in Enceladus' ocean. All authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to H.-W.H. (sean.hsu@lasp.colorado.edu).

METHODS

Dynamics analysis. The dynamical properties derived from nanodust–solar wind interactions (ejection speed: 50–200 km s^{-1}, charge-to-mass ratio: 1,000–20,000 C kg^{-1}, or 2–8 nm in radius assuming +5 V surface potential) links stream particles to an ejection region (ER) at ~8 R_S (R_S is the Saturn radius, 60,268 km) from Saturn[7]. The ER is defined as the region where charged nanoparticles start to gain energy from the co-rotation electric field to escape from the gravity of Saturn. Considering the energy conservation, the ER distribution thus represents the distribution of stream particles' dynamical properties as they are ejected from the system (equation (5) in ref. 7) and reflects the effects of the dust charging process as well as the source location[7]. The ER peak location indicates that their source extends from the inner system to over 8 R_S with strength decreasing outward[7]. It was therefore proposed that stream particles are nanometre-sized Si-rich inclusions in E-ring ice grains released through plasma sputtering erosion[7], as such sputtering is more corrosive on water ice than on Si-bearing minerals[26]. Enceladus is the dominant source of E-ring ice grains[27], suggesting that stream particles also originated from Enceladus.

One major difference between the E ring and other tentative sources (for example, Saturn's main rings and moons) is the vertical extension. E-ring grains can obtain significant inclination because of solar radiation pressure as well as gravitational perturbations from embedded moons[24,28,29]. Nanoparticles released from the E ring would inherit the inclination, as the magnetic field of Saturn aligns well with its rotation axis[30,31]. To examine the proposed hypothesis, we adopt numerical simulations to reconstruct the emission patterns of the Saturnian stream particles, as described below:

The sputtering mass loss rate of E-ring grains. Trajectories of ice grains with initial radius, r, between 0.1 and 5 μm from the dynamics model[29] are used to reconstruct the E-ring profile. E-ring grains follow a power-law size distribution, $n(r) \propto (r/r_0)^{-\mu}$, where r is the grain radius and μ ranges from 4.8 to 5.4 (ref. 24). Weighted with the initial size distribution and normalized to the dust density recorded at the orbit of Enceladus[24], the simulated trajectories are binned to a two-dimensional, axially symmetric dust density map (Extended Data Fig. 1a).

The dust size distribution and the mean dust–plasma relative speed are used to calculate the sputtering mass loss rate of E-ring grains at a given torus segment (ρ, z), expressed by

$$\dot{m}(\rho,z) = \int \Phi_{\mathrm{sput}}(\rho,z) \times A(r,\rho,z) \times m_{H_2O}\,\mathrm{d}r$$

where $A(r,\rho,z)$ is the surface area of E-ring grains with radius r in a given torus segment, ρ and z are the distance to the rotation axis of Saturn and to the ring plane, respectively, and m_{H_2O} is the mass of a water molecule. The sputtering yield (Φ_{sput}) of icy surface in Saturn's magnetosphere is governed by the elastic nuclear collisions from the thermal magnetospheric plasma ions[32–34] and can be written (equation (4) in ref. 33) as:

$$\Phi_{\mathrm{sput}}(\rho,z) = \bar{g}\frac{u(\rho,z) \times n_i(\rho,z)}{4} \times Y(E_i,0),$$

$$\bar{g} = \frac{2}{1-x}\left[1-(\cos\theta_m)^{1-x}\right],$$

$$x \approx 0.3 + 0.13\ln(m_i)$$

where $u(\rho, z)$ is the relative speed between E-ring grains and the magnetospheric plasma ions, $n_i(\rho, z)$ is the plasma ion density, and $m_i = m_{H_2O}$ We use an *ad hoc* plasma model built based on the Cassini measurements[7]. $\theta_m = 80^\circ$ is the ion incident angle beyond which the sputtering yield rapidly decreases[33]. $Y(E_i, 0)$ is the plasma ion sputtering yield of water ice[33,34]. The resulting E-ring mass loss rate is shown in Extended Data Fig. 1b.

Stream particle production rate. This is defined as the amount of escaping nanosilica particles per unit time. Under the assumption that the stream particle production is proportional to the E-ring ice mass loss rate ($\dot{m}$), the production rate (w) is written as:

$$w(r_{sp},\rho,z) = w'(r_{sp},\rho,z) f_{sp} f_{eff}, \quad (1)$$

$$w'(r_{sp},\rho,z) = \frac{\dot{m}(\rho,z)}{m_{sp}(r_{sp})} P_{mass}(r_{sp}) P_{eject}(r_{sp},\rho,z) \quad (2)$$

where f_{sp} is the mass ratio of silica nanoparticle with respect to the water ice in E-ring grains. f_{eff} is the efficiency of nanosilica release via the plasma sputtering process, which depends on the location distribution of nanosilica particles within the ice grain as well as the efficiency of plasma sputtering erosion processes. w', r_{sp}, m_{sp}, P_{mass}, P_{eject} are the normalized production rate, the radius, the mass, the mass distribution function, and the ejection probability of nanosilica stream particles, respectively. Based on the derived size range[7], we assume that stream particles follow a Gaussian distribution with a mean at 4 nm and variance of 2 nm. P_{eject} is calculated from the nanodust ejection model described below. The normalized production rate of 5 nm silica particles is shown in Extended Data Fig. 1c.

Dynamical evolution of charged nanoparticles. The predominant acceleration of charged nanoparticles in Saturn's magnetosphere stems from the outward-pointing co-rotation electric field[5,7,8,30]. In the first order, only positively charged dust particles gain energy and escape. Therefore, the fate of nanoparticles depends on the charging processes, that is, the plasma conditions at the location where they are released. Using the plasma model described previously, the ejection probability map of nanodust particles is simulated. See ref. 7 for the modelling details of the stochastic charging process and the equation of motion of nanoparticles.

Extended Data Fig. 2 shows the P_{eject} maps for 5 nm silica and water ice particles. A successful ejection event is defined as when the required ejection time of a test particle is less than half of its sputtering lifetime. The sputtering yield of water ice is about an order of magnitude larger than that of silicates (for example, $\Phi_{ice} \approx 1.5$ and $\Phi_{silicate} \approx 0.15$ for incident He ions at 500–1,000 eV energy range[26]). We assume that the sputtering lifetime for silica is ten times longer than that of water ice. The particle size decrease due to sputtering is not considered in the simulation[7].

The emission patterns. The dynamical properties of charged test nanoparticles released from the E ring simulated in the above step are converted to ER (equation (5) in ref. 7)) and the latitudinal emission pattern, weighted by the normalized production rate (equation (2)) according to their initial location, as shown in Extended Data Fig. 3a, b. We also modelled the emission patterns assuming that nanosilica particles are ejected directly from Enceladus, to examine our hypothesis (Extended Data Fig. 3c, d).

The nanosilica colloid concentration. f_{sp} in equation (1) can be determined by comparing the modelled stream particle production rate (w) with the CDA stream particle flux measurements, as shown in Fig. 4. The CDA observations are summarized in Extended Data Table 1. We assume that (1) f_{sp} and f_{eff} remain constant through the E-ring grains' lifetime, and (2) nanosilica particles are mixed homogenously in the ice matrix of E-ring grains (that is, $f_{eff} = 1$) so their release is directly proportional to the sputtering erosion rate. Figure 4 shows that the derived f_{sp} ranges from about 150 to 3,900 p.p.m. (parts per million), depending on the adopted E-ring size distribution slope. The conservative lower limit of the dissolved silica concentration at the reaction sites is about 210 p.p.m. (3.5 mM), including the silica solubility at 0 °C (~1 mM, or 60 p.p.m.). This corresponds to minimum reaction temperatures of 250 °C and 120 °C for solution pH values of 9 and 10, respectively (Fig. 2a).

Spectra analysis. *Data set.* The Cassini CDA measures the composition of individual dust grains by time of flight (TOF) mass spectroscopy[9]. Owing to the small mass of stream particles, their impact ionization spectra provide only weak particle mass lines at best[6,7]. In previous investigations[6,7] only Si^+ at 28u (u = unified atomic mass unit) could be identified as a definite particle constituent. Here we aim to go to the absolute detection limit possible with CDA. The goal is to quantify the most prominent elemental stream particle constituent, silicon[6,7], and to identify other elements that are typically abundant in silicate minerals (for example, magnesium or iron). Therefore only spectra with the best particle signals recorded between April 2004 and January 2008 were used for this analysis. The main reason to choose this period is that it provides the highest quality CDA spectra with the lowest possible contamination background. Starting in March 2008, CDA was frequently operating deep inside Enceladus' plume, during which time the refractory constituents of Enceladian ice grains, for example, sodium and potassium salts, might have accumulated and enhanced the CDA target contamination.

From the data set of over 2,000 stream particle spectra, 32 spectra with the highest signal-to-noise ratio of a 28u (±0.6u) signal were selected. A Si^+ signal amplitude of 0.7 μV was chosen as the selection threshold. This value provides clear Si^+ signals as well as a sufficiently large ensemble of spectra. These impact-ionization spectra also show relatively high total ion production (the sum of ions stemming from target material, target contamination and the particle itself). Thus, the selected spectra probably represent the largest detected stream particles at the highest encounter speeds during the observation time.

The selected spectra probably show the highest abundance of particle material (compared to target material and target contamination) and thus provide the highest probability of detecting further elemental particle constituents. Note that even in these spectra, ions from particle compounds only amount to about 1%. To further enhance the signal-to-noise ratio, the spectra were co-added and 'Lee' filtered (Fig. 1). Other exemplary spectra of stream particles can be found elsewhere[6,7].

Spectra interpretation. The selected impacts most probably occurred at speeds above 200 km s^{-1}. In this regime, the energy density is orders of magnitude higher than the molecular bond energies[35,36]. Therefore, similarly to the case of Jovian stream particles[37], only elemental ions are produced upon impact. However, subsequent clustering by collisions of neutral and ionized elements in the impact cloud (before the ions 'feel' the accelerating potential of the TOF spectrometer) can produce two-component ions[37]. In the case of the data set used for this work, this clustering

phenomenon is responsible for the formation of bi-elemental cations from the target material rhodium (Rh_2^+) (Fig. 1). The ratio of Rh^+/Rh_2^+ is about 100. Since rhodium is probably the most abundant constituent of the impact cloud, the intensity of this low-level signature marks the upper limit for the abundance of other non-elemental ions. This also helps to resolve the notorious ambiguity of the 28u signature in mass spectrometry that, besides silicon, can in principle be assigned to cations of N_2, CO, CNH_2 and C_2H_4. Carbon and hydrogen are highly abundant spectrum contaminants from the instrument target, these elements thus cannot be assessed with respect to the composition of stream particles[10]. Although all these components could potentially contribute to the 28u mass line, their abundance can be expected to be very low at most.

From integration of the spectral peaks, abundances and ratios of cationic species in the impact cloud can be directly derived. Ionization probabilities of the different species have to be considered to form conclusion regarding the composition of the particle. This is of particular relevance to reaching a conclusion about the metal to silicon ratio in stream particles, one of the main goals of the spectra analysis. All metals, especially Mg/Ca and Na/K, have a higher probability of forming cations than silicon[38]. The highest possible metal signal in the spectrum shown in Fig. 1 is a peak at an atomic mass of 23–24 u in agreement with sodium (Na^+) and/or magnesium (Mg^+; the adjacent mass lines can not be distinguished here), which is about 5 times less abundant than the Si^+ signature. Two regions with mass lines that can be attributed to K^+/Ca^+ at 39–40 u and unspecified species (at ~3.6 μs, or 56–60 u) are of weak significance, indicating even lower abundances if considered as particle constituents. In contrast to these metals, silicon is not completely converted from elements to cations. It has a higher ionization potential and higher electron affinity, which lead to simultaneous formation of anions, cations and neutrals in the impact cloud. Laboratory calibrations imply the cationization probability of Si to be about 3 times lower than that of Mg[39]. In total, Fig. 1 implies a metal to silicon ratio below 1/10. This ratio of the most metal-depleted silicates is 2/3 and ranges from 1 to 2 for most rock forming minerals. It is possible that traces of Na and K have been transferred to the surface of nanosilica particles from remains of salt-rich carrier ice grains causing the weak signatures at mass 23u and 39u. We note that the observed possible metal signatures are upper limits for the particle constituents, as the CDA target is known to have a low-level contamination of Na and K (ref. 10). A bi-elemental cluster (C_2^+, 24u), formed from the highly abundant carbon contamination, might also significantly contribute to the signal at mass 23–24. Consequently it is possible that the potential weak metal mass lines stem entirely from contamination, and that stream particles are entirely metal-free. To summarize, while we cannot completely rule out that some of the weak signatures have contributions from metal ions stemming from the particle, their abundance is far too low to be in agreement with a rock-forming silicate.

In Fig. 1 the O^+/Si^+ ratio is about 2. However, in contrast to metals, oxygen has a lower probability of forming cations than does silicon. Therefore, the O^+/Si^+ ratio should be clearly below 2 for a pure silica (SiO_2) particle. From laboratory calibration we expect it to be around 1. But as oxygen is known to be a target contaminant that contributes to the O^+ mass line to an unknown extent[10], the observed ratio is consistent with SiO_2.

Stream particle size estimate by integration of the Si^+ signal. By integrating the strongest Si^+ signals, the number of Si^+ ions created by the impinging particle can be calculated. The idea that the impact ionizes all Si atoms is a simplification, but in the case of ultra-fast stream particles it is probably sufficiently accurate to infer a meaningful lower limit for the number of Si atoms in the particle. This in turn allows for mass and size calculation, again a lower limit, if stream particles are assumed to consist solely of SiO_2.

The integrated signal of the Si^+ peak in Fig. 1 is equivalent to about 1,500 ions. As explained above, this signal probably stems from the largest measured stream particles at the highest encounter speed (>200 km s^{-1})[40]. The ions recorded in CDA mass spectra represent about 1/6.5 of the ions that were initially formed[9]. We conclude that the largest stream particles created about 10,000 Si cations upon impact.

Under the assumption of a pure spherical SiO_2 particle, we can now calculate a size from this number. If we want to derive a strict lower limit on the largest particle size, we have to assume a grain built of about 10,000 SiO_2 molecules, which leads to a particle radius of about 6 nm, if we assume a density of 2,200 kg m^{-3} for amorphous silica. As mentioned above, it is highly probable that only a fraction of silicon is converted into cations even at the extreme impact speed of stream particles. A more realistic assumption is that only a third of Si atoms form cations, which gives a maximum particle radius of about 8.5 nm (for comparison, the largest Jovian stream particles reach radii of over 20 nm; refs 8, 37).

Hydrothermal experiments and calculations. We performed hydrothermal experiments based on the methodology and apparatus employed in previous studies[41–43]. The starting mineral powder and solution were introduced into a flexible gold reaction cell, pressurized to 400 bar with a steel alloy autoclave[41]. The pressure condition corresponds to that of Enceladus' rocky core (~150 km below the water–rock interface). The effect of pressure is not critical for estimating the temperature required for nanosilica formation. This is because the silica concentration equilibrated by the serpentine-talc/saponite buffer is not sensitive to pressure range within the core (~100–500 bar)[44]. The flexible gold reaction cell consists of a gold reaction bag and a titanium head[41], which was oxidized before use. The flexible reaction cell allows us to perform an on-line collection of fluid samples during the experiments[41–43]. See ref. 41 for more details.

As starting minerals, we used a mixture of powdered olivine (San Carlos Olivine: $Mg_{1.8}Fe_{0.2}SiO_4$) and orthopyroxene ($MgSiO_3$) (orthopyroxene: olivine = 7: 3 by weight; 15 g in total). These are major minerals known to be abundant in asteroids and comets[45,46]. San Carlos olivine contains trace amounts of spinel and pyroxene, which were the source of Al, Ca and other elements. We synthesized orthopyroxene crystals using the flux method[47,48]. The starting solution (~ 60 g) was an aqueous solution of NH_3 (1.1 mol per kg H_2O) and $NaHCO_3$ (360 mmol per kg H_2O).

We conducted two experiments at different temperatures. One was performed at a constant temperature of 300 °C for ~2,700 h of reaction time. In the other experiment, the temperature was set to 120 °C for an initial ~1,700 h of reaction time, and then increased to 200 °C (~2,300 h of reaction time in total). We measured the concentrations of dissolved silica and other major elements (for example, Na, Mg, Fe, Ca, Al and K) dissolved in the fluid samples during the reaction time with inductively coupled plasma atomic emission spectroscopy (Perkin Elmer). Mineralogical analyses for the solids collected after the experiments were performed with an X-ray diffraction spectrometer (X'PERT-PRO PANalytical). The *in situ* pH of the solution in the experiments was calculated using the measured pH of the fluid samples at room temperature and concentrations of dissolved gas and elements. The *in situ* pH values are calculated as 8.4–8.8, whereas measured pH values at room temperature were 10.1–10.2 at the end of the experiments.

The ΣSiO_2 concentration determined by chemical equilibrium between serpentine and talc/saponite was calculated with equilibrium constants computed with the SUPCRT92 program[49]. Given the similarity in the chemical compositions between talc and saponite, we used the thermodynamic data of talc in the calculations. The solubility of silica at 0 °C was calculated from thermodynamic data of amorphous silica. The concentrations of $HSiO_3^-$ and $NaHSiO_3$(aq.) were calculated for different pH values and at 0.1 mol per kg Na^+ concentration using the equilibrium constants of the following reactions: $SiO_2(aq.) + H_2O \leftrightarrow HSiO_3^- + H^+$ and $HSiO_3^- + Na^+ \leftrightarrow NaHSiO_3(aq.)$.

We observed silica nanoparticle formation by cooling the fluid samples collected in the experiments at 300 °C. A part of the sample was cooled at ~0 °C in an ice bath, and then dialysis treatment (that is, fluid removal) was performed for several minutes. After the dialysis, a drop of the sample was mounted on a slide, and the excess solution was wicked away with tissue paper. Microscopic observations of the slide were performed with a field emission scanning electron microprobe (FE-SEM). Individual and clustered silica nanoparticles were observed (Fig. 2). The typical size of individual particles was ~5–20 nm in diameter. The energy dispersive spectrum indicates that they are composed mainly of Si and O with trace amounts of Na and Ca (Extended Data Fig. 4), which may be adsorbed on the surface of particles.

Timescale of growth of nanosilica in Enceladus' ocean. We estimated this on the basis of the equation shown in the previous study[21]. The primary size of nanosilica formed from alkaline aqueous solution is a few nanometres in radius[12,14–17]. After the formation of these nanosilica particles, the size would increase slowly by precipitation of dissolved silica onto the surface (Ostwald ripening)[12]. The timescale of growth, t_g, of radius from r_s to r by Ostwald ripening in pure water can be described as follows (equation (14) in ref. 21);

$$t_g = r^2 S_0(T_0) / R_0 r_s S_0(T)$$

where R_0 is the dissolution rate of silica, and $S_0(T)$ and $S_0(T_0)$ are the solubility of silica at a given temperature, T, and that at the temperature, T_0, where the experimental data were obtained (25 °C), respectively. According to the previous study[50], R_0 for amorphous silica is 8.8×10^{-15} cm s^{-1} at 0 °C. The ratio of $S_0(T_0)/S_0(T)$ is calculated as 1.67 for $T = 0$°C (ref. 21). Thus, the timescale for a 2-nm-sized nanosilica particle to grow to an 8-nm-sized particle at 0 °C in pure water is estimated as ~20 years. If NaCl is included in the solution, the Ostwald ripening proceeds more rapidly, about 10–100 times faster than in pure water[50]. Thus, the nanosilica particles with radius of ≤8 nm, observed by Cassini CDA, should have been formed within months to years, suggesting continuing hydrothermal activity in Enceladus.

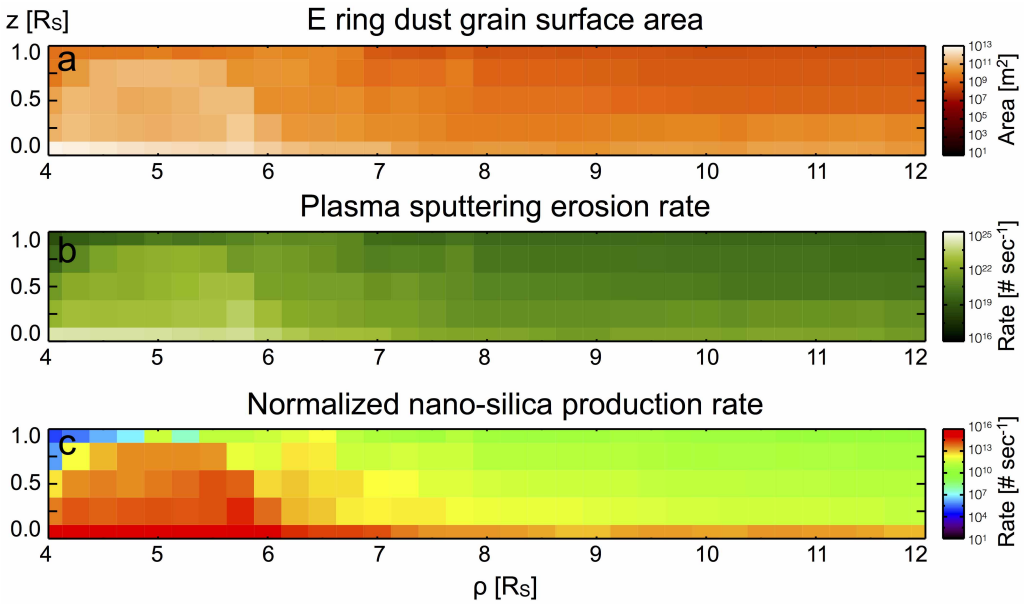
Enceladus' silicon footprint in Saturn's magnetosphere. After being transported to the near-surface plume sources, nanosilica particles eventually become grain inclusions in frozen water droplets[1,2,22] from spray above Enceladus' subsurface liquid plume sources—or they may entrain in the gas flow and serve as condensation seeds in the vent. After entering the E ring they are exposed to Saturn's magnetosphere, separated from ice grains by differential plasma sputtering erosion and eventually ejected into interplanetary space as stream particles.

About 1 mM (60 p.p.m.) of silica might in fact still be dissolved in liquid Enceladus plume sources at 0 °C and become an additional ice grain constituent upon freezing. After sputtering erosion and ionization, this component, as well as erosion of nanosilica particles, contributes to the mass 28 ions observed by the Cassini plasma instruments CAPS (Cassini Plasma Spectrometer)[51] and MIMI (Magnetospheric Imaging Instrument)[52] at different energies.

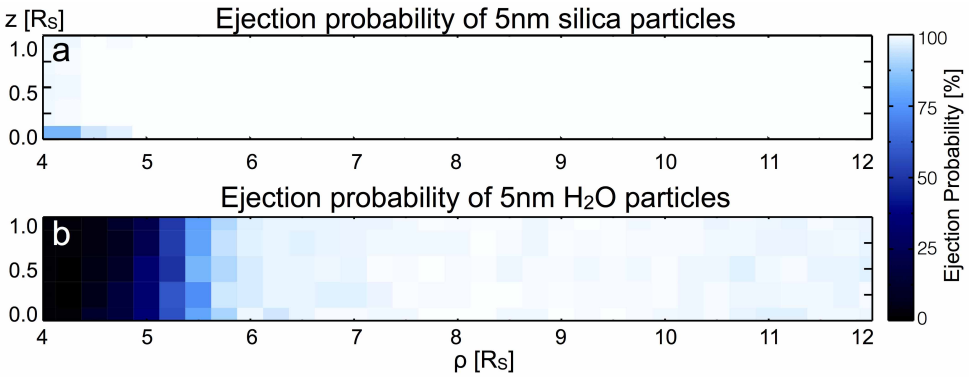
Analysis of CAPS ion measurements[51] shows that the density ratio between the mass 28 and water group ions is about 6×10^{-5}, which corresponds to a mass fraction of ~90 p.p.m. and interestingly is comparable to silica solubility at 0 °C (50 and 120 p.p.m. for pH = 9 and 10, respectively). The mass resolution of Cassini instruments cannot distinguish between $HCNH^+$, CO^+, N_2^+ or Si^+, and therefore no solid conclusion can be drawn for the origin of the mass 28 ions at the current stage[52]. The sputtering yield of Si-water ice mixture is unknown. Nonetheless, the presence of nanosilica particles and ice grains forming from hydrothermal fluids surely will supply the magnetosphere with silicon ions. Future modelling efforts should focus on the ionization, ion lifetime and acceleration processes that may be responsible for the enhanced ratio of $^{28}M^+$ to water-group ions, $(3–7) \times 10^{-3}$, at the 100 keV energy level[52].

Sample size. In data analyses above, no statistical methods were used to predetermine sample size.

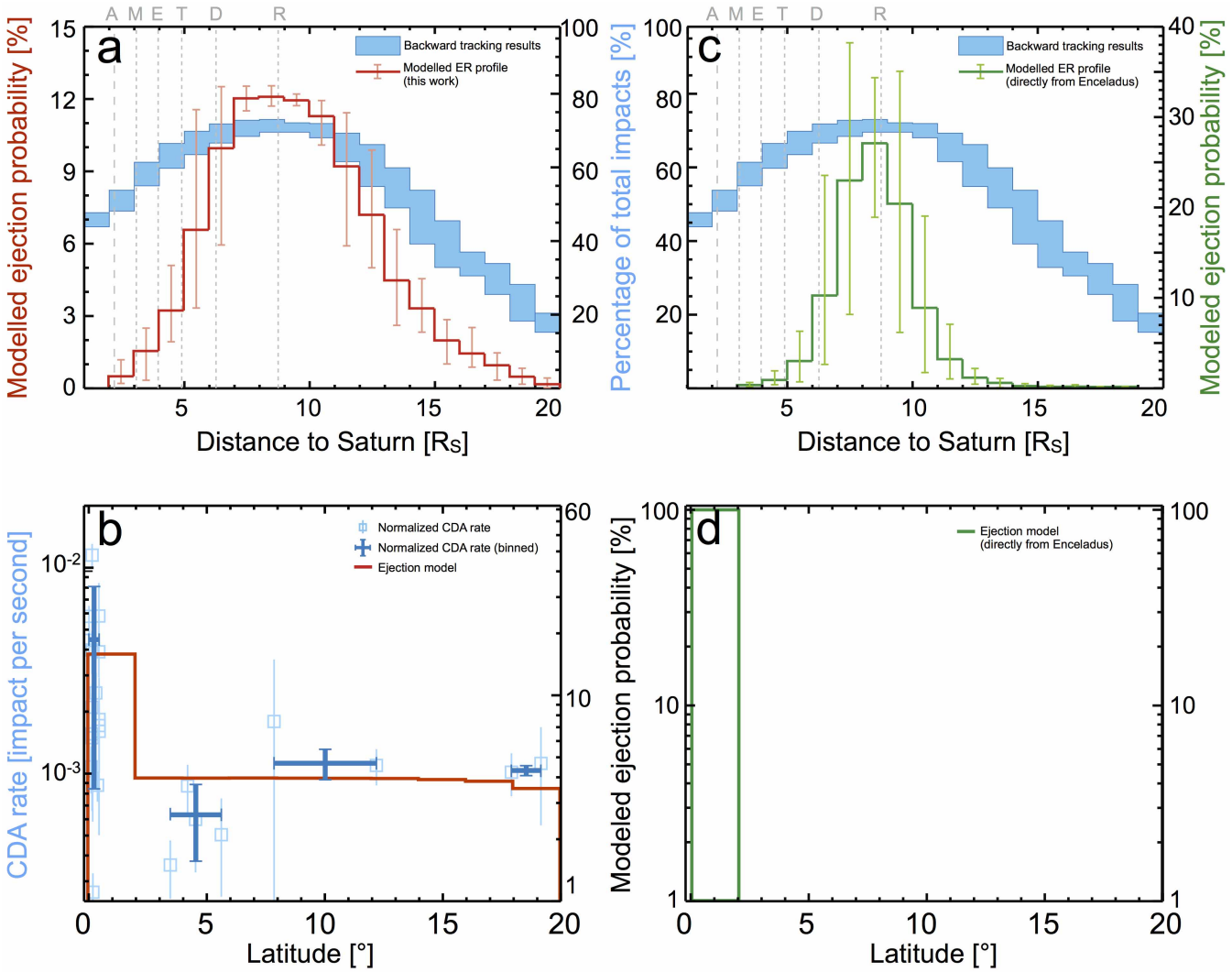
26. Tielens, A. G. G. M. *et al.* The physics of grain-grain collisions and gas-grain sputtering in interstellar shocks. *Astrophys. J.* **431,** 321–340 (1994).
27. Spahn, F. *et al.* Cassini dust measurements at Enceladus and implications for the origin of the E ring. *Science* **311,** 1416–1418 (2006).
28. Horányi, M., Burns, J. & Hamilton, D. G. Dynamics of Saturn's E ring particles. *Icarus* **97,** 248–259 (1992).
29. Horányi, M., Juhász, A. & Morfill, G. E. Large-scale structure of Saturn's E-ring. *Geophys. Res. Lett.* **35,** L04203 (2008).
30. Horányi, M. Dust streams from Jupiter and Saturn. *Phys. Plasmas* **7,** 3847–3850 (2000).
31. Burton, M. E., Dougherty, M. K. & Russell, C. T. Saturn's internal planetary magnetic field. *Geophys. Res. Lett.* **37,** L24105 (2010).
32. Jurac, S., Johnson, R. E. & Richardson, J. D. Saturn's E ring and production of neutral torus. *Icarus* **149,** 384–396 (2001).
33. Johnson, R. E. *et al.* Sputtering of ice grains and icy satellites in Saturn's inner magnetosphere. *Planet. Space Sci.* **56,** 1238–1243 (2008).
34. Shi, M. *et al.* Sputtering of water ice surfaces and the production of extended neutral atmospheres. *J. Geophys. Res.* **100,** 26387–26395 (1995).
35. Hornung, K. & Kissel, J. On shock wave impact ionization of dust particles. *Astron. Astrophys.* **291,** 324–336 (1994).
36. Hornung, K., Malama, Y. & Kestenboim, K. Impact vaporization and ionization of cosmic dust particles. *Astrophys. Space Sci.* **274,** 355–363 (2000).
37. Postberg, F. *et al.* Composition of Jovian dust stream particles. *Icarus* **183,** 122–134 (2006).
38. Stephan, T. TOF-SIMS in cosmochemistry. *Planet. Space Sci.* **49,** 859–906 (2001).
39. Fiege, K. *et al.* Compositional analysis of interstellar dust as seen by the Cassini Cosmic Dust Analyser. In *76th Annual Meteoritical Society Meeting,* http://www.hou.usra.edu/meetings/metsoc2013/pdf/5087.pdf (2013).
40. Hsu, H.-W. *et al.* Probing IMF using nanodust measurements from inside Saturn's magnetosphere. *Geophys. Res. Lett.* **40,** 2902–2906 (2013).
41. Shibuya, T. *et al.* Reactions between basalt and CO_2-rich seawater at 250 and 350°C, 500 bars: implications for the CO_2 sequestration into the modern oceanic crust and composition of hydrothermal vent fluid in the CO_2-rich early ocean. *Chem. Geol.* **359,** 1–9 (2013).
42. Seyfried, W. E. Jr, Foustoukos, D. I. & Fu, Q. Redox evolution and mass transfer during serpentinization: an experimental and theoretical study at 200°C, 500 bar with implications for ultramafic-hosted hydrothermal systems at mid-ocean ridges. *Geochim. Cosmochim. Acta* **71,** 3872–3886 (2007).
43. McCollom, T. M. & Seewald, J. S. Experimental constraints on the hydrothermal reactivity of organic acids and acid anions: I. Formic acid and formate. *Geochim. Cosmochim. Acta* **67,** 3625–3644 (2003).
44. Vance, S. *et al.* Hydrothermal systems in small ocean planets. *Astrobiology* **7,** 987–1005 (2007).
45. Nakamura, T. *et al.* Chondrulelike objects in short-period comet 81P/Wild 2. *Science* **321,** 1664–1667 (2008).
46. Brearley, A. J. in *Meteorites and the Early Solar System II* (eds Lauretta, D. S. & McSween, H. Y.) 587–624 (Univ. Arizona Press, 2006).
47. Ozima, M. Growth of orthoenstatite crystals by the flux method. *J. Jpn Assoc. Mineral. Petrol. Econ. Geol.* **3** (suppl), 97–103 (1982); in Japanese with English abstract.
48. Tachibana, S., Tsuchiyama, A. & Nagahara, H. Experimental study of incongruent evaporation kinetics of enstatite in vacuum and in hydrogen gas. *Geochim. Cosmochim. Acta* **66,** 713–728 (2002).
49. Johnson, J. W., Oelkers, E. H. & Helgeson, H. C. SUPCRT92: a software package for calculating the standard molal thermodynamic properties of minerals, gases, aqueous species, and reactions from 1 to 5000 bar and 0 to 1000°C. *Comput. Geosci.* **18,** 899–947 (1992).
50. Icenhower, J. P. & Dove, P. M. The dissolution kinetics of amorphous silica into sodium chloride solution: effects of temperature and ionic strength. *Geochim. Cosmochim. Acta* **64,** 4193–4203 (2000).
51. Martens, H. R. *et al.* Observations of molecular oxygen ions in Saturn's magnetosphere. *Geophys. Res. Lett.* **35,** L20103 (2008).
52. Christon, S. P. *et al.* Saturn suprathermal O_2^+ and mass-28^+ molecular ions: long-term seasonal and solar variation. *J. Geophys. Res.* **118,** 3446–3463 (2013).
53. Kempf, S. *et al.* The electrostatic potential of E ring particles. *Planet. Space Sci.* **54,** 999–1006 (2006).



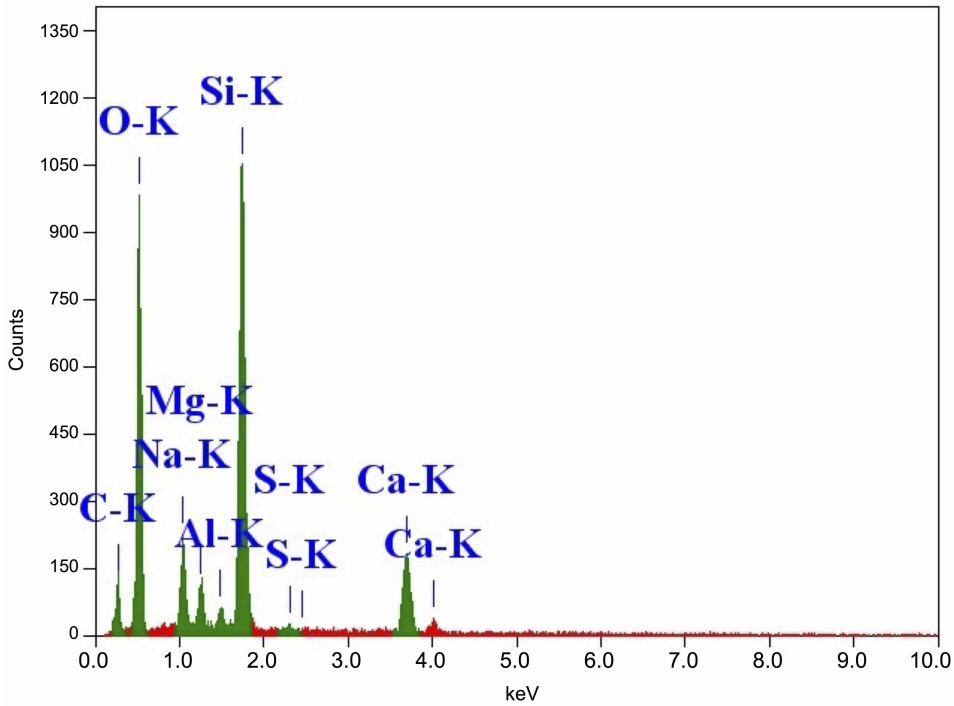
Extended Data Figure 1 | Maps of grain density, sputtering erosion rate, and stream particle production rate in the E-ring region. **a**, The total E-ring ice grain surface area map in the ρ–z frame, where ρ and z are distance to Saturn's rotation axis and to the ring plane, respectively. Note that each bin integrates azimuthally over the entire torus, meaning that the outer bins contain a much larger volume than do the inner ones. **b**, Plasma sputtering erosion rate of E-ring ice grains in torus segments. The total sputtering rate is 8.6×10^{24} H_2O molecules per second, lower but still comparable to the 4.5×10^{25} H_2O molecules per second derived in ref. 32. **c**, Normalized nanoparticle production rate in particles per second. R_S, Saturn radius.



Extended Data Figure 2 | Ejection probability of 5-nm particles from the E ring. **a**, For silica nanoparticles, the ejection probability is mostly close to unity (except within $4.5R_S$). The higher local plasma density there leads to negative dust potential and thus reduces the ejection probability[7]. The typical timescale for silica nanoparticles to acquire sufficient kinetic energy to escape is of the order of a day[7]. **b**, Water ice nanoparticles have lower secondary emission and are charged less positively and thus are less likely to be ejected. This 'forbidden region' (the black region) extends further outward to $\sim 5.5\,R_S$, consistent with the CDA measurements[53].



Extended Data Figure 3 | Stream particle emission patterns. **a**, Ejection region (ER) profiles, derived from the nanodust and solar wind measurements (blue)[7] and the ejection model (red), both peak at 7–9R_S. The uncertainty of both profiles stems from the adopted co-rotation fraction of Saturn's magnetosphere (80–100%), which determines the electromagnetic acceleration amplitude. The location of the outer rim of Saturn's A ring and the orbits of icy satellites are marked by grey dashed lines. **b**, Latitudinal-dependent ejection pattern. Scatter and binned stream particle rates (normalized to 25R_S distance) are shown in blue squares and crosses, respectively. The vertical length of the crosses represents the standard deviation of the stream particle rate in the corresponding bin. Our model (red) reproduces the measured trend. **c**, **d**, Modelled patterns assuming direct ejection from Enceladus. While the ER profile is similar, these particles are only ejected along the ring plane.



Extended Data Figure 4 | Energy dispersive spectrum of clustered silica nanoparticles formed from the fluid sample. See Methods for details.

Extended Data Table 1 | Stream particle flux measurements

UTC (medium)	Dur. (minute)	Dist. (R_S)	Lat. (°)	Impacts	Impact rate (# second^{-1})	Production rate (# second^{-1})
2004-346T01:05	280	36.4	-5.62	4	2.4×10^{-4}	8.6×10^{17}
2004-347T13:40	579	28.0	-3.45	10	2.9×10^{-4}	1.2×10^{17}
2004-348T22:25	585	18.0	0.17	18	5.1×10^{-4}	9.1×10^{16}
2005-063T00:07	541	38.2	-0.10	15	4.6×10^{-4}	3.7×10^{17}
2005-084T18:55	135	33.5	-0.14	18	2.2×10^{-3}	1.4×10^{18}
2005-084T23:25	204	32.8	-0.14	15	1.2×10^{-3}	7.2×10^{17}
2005-085T23:25	187	28.1	-0.17	8	7.1×10^{-4}	3.1×10^{17}
2005-099T05:15	564	36.3	-4.19	14	4.1×10^{-4}	3.0×10^{17}
2005-100T00:00	264	34.5	-4.52	5	3.1×10^{-4}	2.1×10^{17}
2005-228T16:05	464	31.4	-17.9	18	6.5×10^{-4}	1.9×10^{18}
2005-282T10:50	114	25.8	-0.30	16	2.3×10^{-3}	8.4×10^{17}
2005-330T11:50	204	13.9	-0.37	35	2.8×10^{-3}	3.0×10^{17}
2005-336T21:45	319	39.0	-0.02	35	1.8×10^{-3}	1.5×10^{18}
2006-116T17:18	187	24.2	-0.14	18	1.6×10^{-3}	5.1×10^{17}
2006-136T05:23	86	40.9	0.14	3	5.8×10^{-4}	5.3×10^{17}
2006-146T17:20	150	35.9	0.44	7	7.8×10^{-4}	5.5×10^{17}
2006-147T12:50	50	40.1	0.43	2	6.7×10^{-4}	5.9×10^{17}
2006-147T16:55	191	40.6	0.43	8	6.9×10^{-4}	6.3×10^{17}
2006-261T04:35	858	37.6	12.2	25	4.9×10^{-4}	2.0×10^{18}
2006-308T04:35	75	28.1	19.2	4	8.9×10^{-4}	2.1×10^{18}

Twenty observations obtained when Saturn was within 28° of the CDA bore-sight were selected. 278 impacts were registered during the total 100.8 h observation time. Data showing the flux enhancement caused by solar wind–nanodust interactions[8,40] were excluded. The impact rate is normalized to a Saturn distance of 25 R_S (inverse-square law) and is converted to production rate by the modelled flux–latitude relation (Extended Data Fig. 3b). The weighted production rate is $(8.3 \pm 6.3) \times 10^{17}$ particles per second, corresponding to 1.0 ± 0.7 g per second (assuming a mean particle radius of 5 nm). UTC (medium), medium time of observation in Coordinated Universal Time; Dur., duration; Dist., distance; Lat, latitude.

LETTER

doi:10.1038/nature14223

Observation of antiferromagnetic correlations in the Hubbard model with ultracold atoms

Russell A. Hart[1]*, Pedro M. Duarte[1]*, Tsung-Lin Yang[1], Xinxing Liu[1], Thereza Paiva[2], Ehsan Khatami[3], Richard T. Scalettar[4], Nandini Trivedi[5], David A. Huse[6] & Randall G. Hulet[1]

Ultracold atoms in optical lattices have great potential to contribute to a better understanding of some of the most important issues in many-body physics, such as high-temperature superconductivity[1]. The Hubbard model—a simplified representation of fermions moving on a periodic lattice—is thought to describe the essential details of copper oxide superconductivity[2]. This model describes many of the features shared by the copper oxides, including an interaction-driven Mott insulating state and an antiferromagnetic (AFM) state. Optical lattices filled with a two-spin-component Fermi gas of ultracold atoms can faithfully realize the Hubbard model with readily tunable parameters, and thus provide a platform for the systematic exploration of its phase diagram[3,4]. Realization of strongly correlated phases, however, has been hindered by the need to cool the atoms to temperatures as low as the magnetic exchange energy, and also by the lack of reliable thermometry[5]. Here we demonstrate spin-sensitive Bragg scattering of light to measure AFM spin correlations in a realization of the three-dimensional Hubbard model at temperatures down to 1.4 times that of the AFM phase transition. This temperature regime is beyond the range of validity of a simple high-temperature series expansion, which brings our experiment close to the limit of the capabilities of current numerical techniques, particularly at metallic densities. We reach these low temperatures using a compensated optical lattice technique[6], in which the confinement of each lattice beam is compensated by a blue-detuned laser beam. The temperature of the atoms in the lattice is deduced by comparing the light scattering to determinant quantum Monte Carlo simulations[7] and numerical linked-cluster expansion[8] calculations. Further refinement of the compensated lattice may produce even lower temperatures which, along with light scattering thermometry, would open avenues for producing and characterizing other novel quantum states of matter, such as the pseudogap regime and correlated metallic states of the two-dimensional Hubbard model.

A two-spin-component Fermi gas in a simple cubic optical lattice may be described by a single-band Hubbard model with nearest-neighbour tunnelling t and on-site interaction $U>0$. At a density n of one atom per site, and for sufficiently large U/t, there is a crossover from a 'metallic' state to a Mott insulating regime[9] as the temperature T is reduced below U. The Mott regime has been demonstrated with ultracold atoms in an optical lattice by observing the reduction of doubly occupied sites[10] and the related reduction of the global compressibility[11]. For T below the Néel ordering temperature T_N, which for $U \gg t$ is approximately equal to the exchange energy $J=4t^2/U$, the system undergoes a phase transition to an AFM state[12]. In the context of quantum simulations, AFM phases of Ising spins have been previously engineered with bosonic atoms in an optical lattice[13] and with spin-1/2 ions[14,15]. Also, nearest-neighbour AFM correlations due to magnetic exchange have been observed along one dimension of an anisotropic lattice[16]. The same experiment achieved temperatures as low as $T=0.95t\approx 2.6T_N$ when the lattice was configured to be isotropic[17], where $T_N=0.36t$ is the maximal value of the Néel transition temperature[12,18,19].

Our experiments are performed with an all-optically produced[20], quantum degenerate, two-state mixture of the two lowest hyperfine ground states of fermionic ^{6}Li atoms, which we label $|\uparrow\rangle$ and $|\downarrow\rangle$. The repulsive interaction between atoms in states $|\uparrow\rangle$ and $|\downarrow\rangle$ is controlled via a magnetic Feshbach resonance[21], which we use to set the s-wave scattering length a_s in the range from $80a_0$ to $560a_0$, where a_0 is the Bohr radius. A simple cubic optical lattice is formed at the intersection of three mutually perpendicular infrared retroreflected laser beams. We can dynamically rotate the polarization of the retroreflection, and thus continually adjust the potential between a lattice and a harmonic dimple trap. The overall confinement produced by the Gaussian envelope of each infrared lattice beam is partially compensated with a superimposed, non-retroreflected, blue-detuned laser beam[6,22]. The compensation beams serve three purposes: (1) they help flatten the confining potential in order to enlarge the volume of the AFM phase; (2) they provide a way to maintain the central density near $n\approx 1$ as the lattice is loaded; and (3) they may mitigate the effects of heating in the lattice by lowering the threshold for evaporation.

A degenerate sample with total atom number N between 1.0×10^5 and 2.5×10^5 is prepared in the harmonic dimple trap (without compensation) at a temperature $T/T_F=0.04\pm0.02$, where T_F is the Fermi temperature. The lattice is turned on slowly to a central depth of $v_0=7E_r$ (see Methods), where $E_r=h^2/(2m\lambda^2)$ is the recoil energy, h is Planck's constant, m is the atomic mass, and $\lambda=1{,}064$ nm is the wavelength of the lattice beams. While loading the lattice, the intensities of the compensation beams are adjusted to maintain a peak density $n\approx 1$. We have measured the temperature in the dimple trap before and after transferring the atoms to the lattice (see Methods and Extended Data Fig. 3), and have observed that the compensating beams mitigate heating in the lattice, perhaps by allowing continued evaporative cooling[6] or by a reduction of three-body loss.

Bragg scattering of near-resonant light[23–25] is depicted in Fig. 1. The Bragg condition for scattering from an AFM-ordered sample is satisfied when the momentum $\boldsymbol{Q}$ transferred to a scattered photon is equal to $\boldsymbol{\pi}$, where $\boldsymbol{\pi}=(2\pi/a)(-1/2,-1/2,1/2)$ is a reciprocal lattice vector of the magnetic sublattice, and $a=\lambda/2$ is the lattice spacing. Cameras are positioned to detect scattering at $\boldsymbol{Q}=\boldsymbol{\pi}$ and also at $\boldsymbol{Q}=\boldsymbol{\theta}$, a momentum transfer that does not satisfy the Bragg condition and is used as a control. We obtain spin sensitivity, in analogy to neutron scattering in condensed matter, by setting the Bragg laser frequency between the optical transition frequencies for the two spin states[26,27]. Prior to the measurement, we jump v_0 to $20E_r$ in a few microseconds to lock the atoms in place (see Methods), and then illuminate them *in situ* for 1.7 μs with

[1]Department of Physics and Astronomy and Rice Quantum Institute, Rice University, 6100 Main Street, Houston, Texas 77005, USA. [2]Instituto de Fisica, Universidade Federal do Rio de Janeiro, Caixa Postal 68.528, Rio de Janeiro RJ, 21941-972, Brazil. [3]Department of Physics and Astronomy, San Jose State University, 1 Washington Square, San Jose, California 95192, USA. [4]Department of Physics, University of California, 1 Shields Avenue, Davis, California 95616, USA. [5]Department of Physics, The Ohio State University, 191 West Woodruff Avenue, Columbus, Ohio 43210, USA. [6]Department of Physics, Princeton University, Princeton, New Jersey 08544, USA.
*These authors contributed equally to this work.

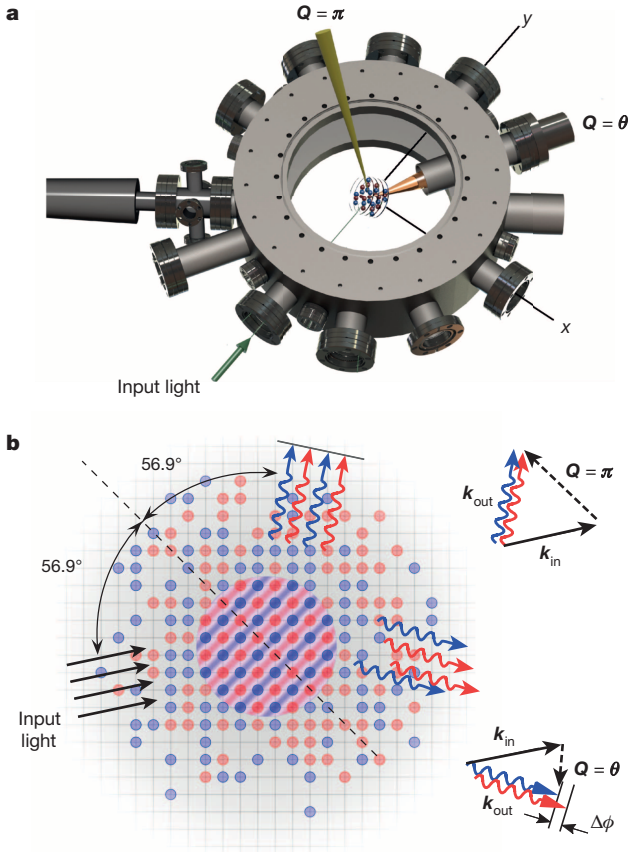


Figure 1 | Schematic depiction of Bragg scattering. **a**, Rendering of the experimental set-up used for Bragg scattering. Light is collected for momentum transfers $\boldsymbol{Q}=\boldsymbol{\pi}$ and $\boldsymbol{Q}=\boldsymbol{\theta}$. A bias magnetic field, which sets the quantization axis and the interaction strength, points in the z direction. The input Bragg beam lies in the y–z plane, and its wavevector makes an angle of 3° with the positive y axis. **b**, The two spin states are denoted by red and blue circles. AFM order develops at the Mott plateau, shown here to be located in the centre, where $n \approx 1$. AFM correlations are suppressed outside the central region where $n < 1$. Bragg scattering requires the input and output wavevectors, $\boldsymbol{k}_{\rm in}$ and $\boldsymbol{k}_{\rm out}$, respectively, to satisfy the Bragg condition $\boldsymbol{k}_{\rm out}-\boldsymbol{k}_{\rm in}=\boldsymbol{\pi}$. The red and blue arrows denote light scattered from one spin state or the other. The two spin states scatter with opposite phase shifts, so that their respective sublattices interfere constructively for $\boldsymbol{Q}=\boldsymbol{\pi}$. For a different momentum transfer $\boldsymbol{k}_{\rm out}-\boldsymbol{k}_{\rm in}=\boldsymbol{\theta}$, scattering is relatively insensitive to AFM correlations owing to the lack of constructive interference between the scattered photons, which have random relative phases $\Delta\phi$.

the Bragg probe. Alternatively, we can suddenly turn off the $20E_{\rm r}$ lattice and illuminate the atoms after time-of-flight τ.

Figure 2 shows the results of simultaneous measurements of the scattered intensity for $\boldsymbol{Q}=\boldsymbol{\pi}$ and $\boldsymbol{Q}=\boldsymbol{\theta}$ ($I_{\boldsymbol{\pi}}$ and $I_{\boldsymbol{\theta}}$, respectively), as a function of τ. After a few microseconds of expansion, when the extent of the atomic wave packets becomes comparable to the lattice spacing, the light scattered from correlated spins no longer interferes constructively at the detector. More precisely, the Debye–Waller factor $e^{-2W_{\boldsymbol{Q}}(\tau)} = \exp\left[-\sum_{i=x,y,z}\boldsymbol{Q}_i^2\left\langle r_i^2\right\rangle_\tau\right]$ decays to zero after a sufficiently long τ (see Methods) and the sample is effectively uncorrelated. Here r_i is the displacement of an atom from the centre of the lattice site at which it was initially localized.

By comparing the intensity of the light scattered *in situ* ($\tau = 0$) to that after sufficiently long τ ($I_{\boldsymbol{Q}0}$ and $I_{\boldsymbol{Q}\infty}$, respectively), we effectively normalize the Bragg scattering signal to the diffuse scattering background of an uncorrelated sample, achieving high sensitivity to magnetic ordering and strong rejection of common mode systematics. Figure 2 shows that there is enhanced scattering at $\tau = 0$ relative to the uncorrelated cloud ($\tau = 9\,\mu$s) for $\boldsymbol{Q}=\boldsymbol{\pi}$, whereas for $\boldsymbol{Q}=\boldsymbol{\theta}$ scattering at

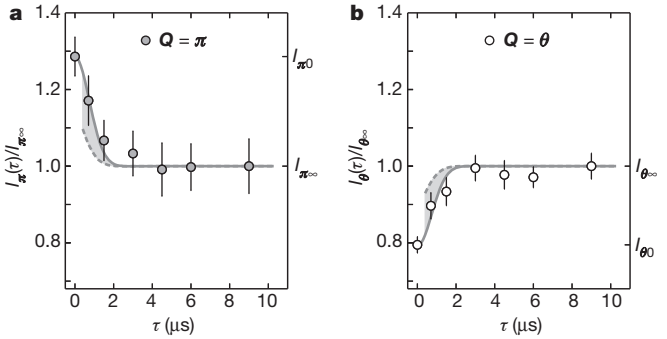


Figure 2 | Time-of-flight measurement of scattered intensity from a sample with AFM correlations. **a**, Normalized intensity of Bragg-scattered light ($\boldsymbol{Q}=\boldsymbol{\pi}$) as a function of time-of-flight τ. The *in situ* ($\tau = 0$) scattered intensity is denoted $I_{\boldsymbol{Q}0}$, while the intensity after sufficiently long τ, corresponding to an effectively uncorrelated sample, is denoted $I_{\boldsymbol{Q}\infty}$. **b**, For $\boldsymbol{Q}=\boldsymbol{\theta}$ the *in situ* sample shows a reduction of scattering, as compared to long τ, due to the presence of double occupancies and to the presence of AFM spin correlations (see text). Each data point and error bar is the mean and standard error of the mean (s.e.m.) of at least 17 measurements of the scattered intensity. The solid grey line is the intensity calculated using the value of the Debye–Waller factor at τ, whereas the dashed grey line uses the average value of the Debye–Waller factor during the 1.7 μs exposure of the Bragg probe (see text and Methods).

$\tau = 0$ is reduced, such that $I_{\boldsymbol{\theta}0}/I_{\boldsymbol{\theta}\infty} < 1$. Double occupancies, present as 'virtual' states even at low temperatures[28], reduce coherent scattering in all directions, since each atom in the pair has opposite spin and therefore scatters with opposite phase. For $\boldsymbol{Q}=\boldsymbol{\pi}$ the coherent enhancement from AFM spin correlations exceeds this reduction. Furthermore, the coherent enhancement of the signal along $\boldsymbol{Q}=\boldsymbol{\pi}$ suppresses the scattered intensity in other directions.

For a momentum transfer $\boldsymbol{Q}$, the spin structure factor $S_{\boldsymbol{Q}}$ of the sample is defined as

$$S_{\boldsymbol{Q}} \equiv \frac{4}{N}\sum_{i,j} e^{i\boldsymbol{Q}\cdot\left(\boldsymbol{R}_i-\boldsymbol{R}_j\right)}\left\langle\sigma_{zi}\sigma_{zj}\right\rangle \qquad (1)$$

Here N is the total number of atoms, the sums extend over all lattice sites i and j, $\boldsymbol{R}_j$ is the location of the jth site, and σ_{zj} is the z component of the spin operator for the jth site:

$$\sigma_{zj}|0\rangle_j = 0|0\rangle_j,\ \sigma_{zj}|\uparrow\rangle_j = +\frac{1}{2}|\uparrow\rangle_j,\ \sigma_{zj}|\downarrow\rangle_j = -\frac{1}{2}|\downarrow\rangle_j,\ \sigma_{zj}|\uparrow\downarrow\rangle_j = 0|\uparrow\downarrow\rangle_j$$

In a sample with complete AFM ordering $S_{\boldsymbol{\pi}} \approx N$, whereas for uncorrelated samples in the lattice $S_{\boldsymbol{\pi}} \leq 1$ and $S_{\boldsymbol{\theta}} \leq 1$. The choice of the z spin component for this analysis is arbitrary, as each of the other axes would result in the same value for $S_{\boldsymbol{Q}}$ in the absence of a symmetry-breaking field. In the limit of tightly localized wavefunctions ($e^{-2W_{\boldsymbol{Q}}(\tau=0)} \approx 1$), and for a weak probe, the spin structure factor is $S_{\boldsymbol{Q}} \approx I_{\boldsymbol{Q}0}/I_{\boldsymbol{Q}\infty}$. We determine the spin structure factor by measuring the scattered intensities $I_{\boldsymbol{Q}0}$ and $I_{\boldsymbol{Q}\infty}$ and applying a correction to account for the *in situ* Debye–Waller factor in the $20E_{\rm r}$ lattice and for saturation of the atomic transition, which generates a small component of inelastically scattered light (see Methods).

Within the local density approximation (LDA) we model the sample by considering each point in the trap as a homogeneous system in equilibrium at a temperature T, with local values of the chemical potential and the Hubbard parameters determined by the trap potential. The spin structure factor of the sample $S_{\boldsymbol{Q}}$ can then be expressed as the integral over the trap of the local spin structure factor per lattice site, $s_{\boldsymbol{Q}}$. Figure 3a shows numerical calculations of $s_{\boldsymbol{\pi}}$ for various temperatures in a homogeneous lattice with $U/t = 8$, close to where $T_{\rm N}$ is maximal[12]. The figure shows that $s_{\boldsymbol{\pi}}$ is sharply peaked around $n = 1$ and grows rapidly as T approaches $T_{\rm N}$ from above.

Figure 3b and c shows n and $s_{\boldsymbol{\pi}}$ profiles, respectively, calculated for our experimental parameters at various values of U_0/t_0, where U_0 and

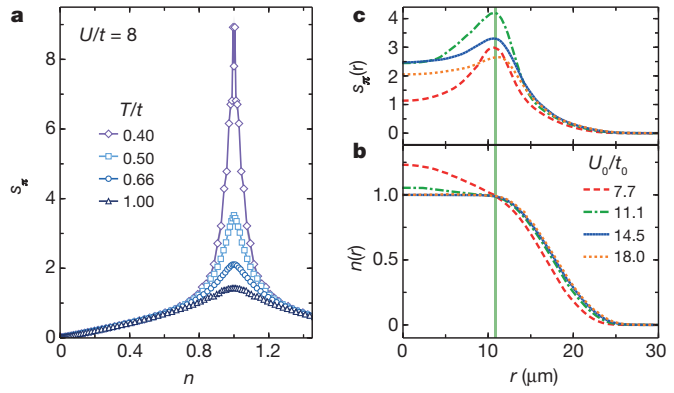


Figure 3 | Numerical calculations. **a**, Spin structure factor per lattice site s_π as a function of n in a homogeneous lattice for several temperatures (see Methods). s_π is sharply peaked near $n = 1$ and diverges as T approaches T_N. **b**, Density profiles calculated at $T/t_0 = 0.6$ for different U_0/t_0, using in each case the value of N that maximizes the experimentally measured S_π (see text and Extended Data Fig. 2). **c**, Profiles of the local spin structure factor $s_\pi(r)$, for the same conditions as in **b**. The vertical green line in **b** and **c** marks the radius at which $s_\pi(r)$ is maximized for $U_0/t_0 = 11.1$ (see text).

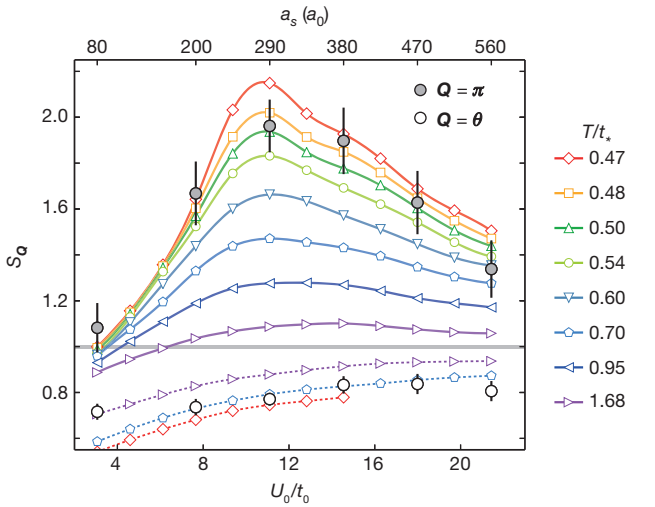


Figure 4 | Spin structure factor. Measured S_π (filled circles) and $S_{\boldsymbol{\theta}}$ (open circles) at optimized N (see text) for various U_0/t_0. The values of the s-wave scattering length corresponding to U_0/t_0 for the experimental points are shown along the top axis. For each point at least 40 *in situ* and 40 time-of-flight measurements of the scattered intensities are used to obtain the spin structure factor. Error bars are obtained from the s.e.m. of the scattered intensities; the raw data are presented in Extended Data Fig. 5. Numerical calculations of S_π (open symbols, lines as guide to the eye) and $S_{\boldsymbol{\theta}}$ (open symbols, dashed lines as guide to the eye) are shown for various values of T/t_*. The numerical calculations for $S_{\boldsymbol{\theta}}$ are unreliable for $T/t_* < 0.7$ and $U_0/t_0 > 15$. $S_{\boldsymbol{\theta}}$ decreases slightly for weak interactions, where the fraction of double occupancies increases.

t_0 denote the local values of the Hubbard parameters at the centre of the trap. As seen in Fig. 3b, only a fraction of the atoms in the sample is near $n = 1$, where AFM correlations are maximal. The finite extent of the lattice beams causes the lattice depth to decrease with distance from the centre, resulting in an increasing t such that both U/t and T/t decrease with increasing radius for constant T (see Extended Data Fig. 1). The radial decrease in T/t causes $s_\pi(r)$ to maximize at the largest radius for which the density is $n \approx 1$. For large U_0/t_0 the cloud exhibits an $n = 1$ Mott plateau and $s_\pi(r)$ is maximized at the outermost radius of the plateau.

In the experiment, we measure $S_{\boldsymbol{Q}}$ as a function of U_0/t_0. At each value of U_0/t_0 we vary the atom number N to maximize the measured S_π (see Methods and Extended Data Fig. 2). According to the picture presented above, this has the effect of optimizing the size and location of the $n = 1$ region of the cloud such that AFM correlations are maximized. The compensation strength g_0, which is the same for all U_0/t_0, was also adjusted to maximize S_π. We found the optimum to be $g_0 = 3.7E_r$ at a lattice depth $v_0 = 7E_r$ (see Methods). Besides the equilibrium considerations regarding the optimal size and location of the Mott plateau, we believe that the dynamical adjustment of g_0 during the lattice turn-on reduces the time for the system to equilibrate, by minimizing the deviation of the equilibrium density distribution in the final potential from the starting density distribution in the dimple trap before loading the lattice.

Figure 4 shows the measured values of S_π and $S_{\boldsymbol{\theta}}$ at optimal N for various values of U_0/t_0 (see Extended Data Fig. 5 for the raw counts at the CCD cameras). We find that S_π is peaked for $11 < U_0/t_0 < 15$. In contrast, the measurements of $S_{\boldsymbol{\theta}}$ vary little over the range of interaction strengths, consistent with an absence of coherent Bragg scattering in this direction. Measurements of S_π after hold time in the lattice show that the Bragg signal decays for larger temperatures (see Extended Data Fig. 4). Comparing the measured S_π with numerical calculations for a homogeneous lattice (for example, those in Fig. 3a) allows us to set a trap-independent upper limit on the temperature, which we determine to be $T/t_0 < 0.7$.

Precise thermometry is obtained by comparing the measured S_π with numerical calculations averaged over the trap density distribution for different values of T. The results of such numerical calculations are shown in Fig. 4, labelled by the value of T/t_*, which we define as the local value of T/t at the radius where the spin structure factor per lattice site is maximal (see Fig. 3c). At $U_0/t_0 = 11.1$, where measured AFM correlations are maximal, we find $T/t_* = 0.51 \pm 0.06$, where the uncertainty is due to the statistical error in the measured S_π and the systematic uncertainty in the lattice parameters used for the numerical calculation. This temperature is consistent with the data at all values of U_0/t_0. We warn, however, that for values of $U/t > 10$ a single-band Hubbard model may not be adequate, as corrections involving higher bands may become non-negligible[27,29].

As was shown in Fig. 3b, for $U_0/t_0 = 11.1$ the dominant contribution to S_π comes from the outermost radius of the Mott plateau. At that radius, the local value of U/t is $U_*/t_* = 9.1$, consistent with determinant quantum Monte Carlo (DQMC) calculations for the homogeneous lattice[12,18,19], which find T_N to be maximized for U/t between 8 and 9. For $U_0/t_0 = 11.1$, $t_* = 1.3$ kHz, so we can infer the temperature of the system to be $T = 32 \pm 4$ nK. In terms of T_N, the temperature is $T/T_N = 1.42 \pm 0.16$. At this temperature, the numerical calculations indicate that the correlation length is approximately the lattice spacing. The calculations show that the entropy per particle in the trap is $S/(Nk_B) \approx 0.76$, where k_B is the Boltzmann constant (see Extended Data Fig. 6). This entropy range is consistent with T/T_F measured in the harmonic dimple trap[30] after a lattice round trip, as shown in Extended Data Fig. 3.

We have observed AFM correlations in the three-dimensional (3D) Hubbard model using ultracold atoms in an optical lattice via spin-sensitive Bragg scattering of light. Because magnetic order is extremely sensitive to T in the vicinity of T_N, Bragg scattering provides precise thermometry in regimes previously inaccessible to quantitative temperature measurements. Whereas previous cold-atom experiments on the 3D Fermi–Hubbard model were in a temperature regime that could be accurately represented by a simple high-temperature series expansion, the data presented here are near the limit of the capabilities of advanced numerical simulations. Our experimental set-up can be configured to study the two-dimensional (2D) Hubbard model in an array of planes; further progress to lower temperature will put us in a position to answer questions about competing pairing mechanisms in 2D, and may ultimately resolve the long-standing question of d-wave superconductivity in the Hubbard model.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 22 July; accepted 4 December 2014.
Published online 23 February 2015.

1. Hofstetter, W., Cirac, J. I., Zoller, P., Demler, E. & Lukin, M. D. High-temperature superfluidity of fermionic atoms in optical lattices. *Phys. Rev. Lett.* **89,** 220407 (2002).
2. Anderson, P. W. The resonating valence bond state in La_2CuO_4 and superconductivity. *Science* **235,** 1196–1198 (1987).
3. Jaksch, D. & Zoller, P. The cold atom Hubbard toolbox. *Ann. Phys.* **315** (spec. issue), 52–79 (2005).
4. Bloch, I., Dalibard, J. & Zwerger, W. Many-body physics with ultracold gases. *Rev. Mod. Phys.* **80,** 885–964 (2008).
5. McKay, D. C. & DeMarco, B. Cooling in strongly correlated optical lattices: prospects and challenges. *Rep. Prog. Phys.* **74,** 054401 (2011).
6. Mathy, C. J. M., Huse, D. A. & Hulet, R. G. Enlarging and cooling the Néel state in an optical lattice. *Phys. Rev. A* **86,** 023606 (2012).
7. Blankenbecler, R., Scalapino, D. J. & Sugar, R. L. Monte Carlo calculations of coupled boson-fermion systems. I. *Phys. Rev. D* **24,** 2278–2286 (1981).
8. Rigol, M., Bryant, T. & Singh, R. R. P. Numerical linked-cluster approach to quantum lattice models. *Phys. Rev. Lett.* **97,** 187202 (2006).
9. Imada, M., Fujimori, A. & Tokura, Y. Metal-insulator transitions. *Rev. Mod. Phys.* **70,** 1039–1263 (1998).
10. Jördens, R., Strohmaier, N., Günter, K., Moritz, H. & Esslinger, T. A Mott insulator of fermionic atoms in an optical lattice. *Nature* **455,** 204–207 (2008).
11. Schneider, U. *et al.* Metallic and insulating phases of repulsively interacting fermions in a 3D optical lattice. *Science* **322,** 1520–1525 (2008).
12. Staudt, R., Dzierzawa, M. & Muramatsu, A. Phase diagram of the three-dimensional Hubbard model at half filling. *Eur. Phys. J. B* **17,** 411–415 (2000).
13. Simon, J. *et al.* Quantum simulation of antiferromagnetic spin chains in an optical lattice. *Nature* **472,** 307–312 (2011).
14. Kim, K. *et al.* Quantum simulation of frustrated Ising spins with trapped ions. *Nature* **465,** 590–593 (2010).
15. Britton, J. W. *et al.* Engineered two-dimensional Ising interactions in a trapped-ion quantum simulator with hundreds of spins. *Nature* **484,** 489–492 (2012).
16. Greif, D., Uehlinger, T., Jotzu, G., Tarruell, L. & Esslinger, T. Short-range quantum magnetism of ultracold fermions in an optical lattice. *Science* **340,** 1307–1310 (2013).
17. Imriška, J. *et al.* Thermodynamics and magnetic properties of the anisotropic 3D Hubbard model. *Phys. Rev. Lett.* **112,** 115301 (2014).
18. Paiva, T., Loh, Y. L., Randeria, M., Scalettar, R. T. & Trivedi, N. Fermions in 3D optical lattices: cooling protocol to obtain antiferromagnetism. *Phys. Rev. Lett.* **107,** 086401 (2011).
19. Kozik, E., Burovski, E., Scarola, V. W. & Troyer, M. Néel temperature and thermodynamics of the half-filled three-dimensional Hubbard model by diagrammatic determinant Monte Carlo. *Phys. Rev. B* **87,** 205102 (2013).
20. Duarte, P. M. *et al.* All-optical production of a lithium quantum gas using narrow-line laser cooling. *Phys. Rev. A* **84,** 061406 (2011).
21. Houbiers, M., Stoof, H. T. C., McAlexander, W. I. & Hulet, R. G. Elastic and inelastic collisions of ^{6}Li atoms in magnetic and optical traps. *Phys. Rev. A* **57,** R1497–R1500 (1998).
22. Ma, P. N. *et al.* Influence of the trap shape on the detection of the superfluid-Mott-insulator transition. *Phys. Rev. A* **78,** 023605 (2008).
23. Birkl, G., Gatzke, M., Deutsch, I. H., Rolston, S. L. & Phillips, W. D. Bragg scattering from atoms in optical lattices. *Phys. Rev. Lett.* **75,** 2823–2826 (1995).
24. Weidemüller, M., Görlitz, A., Hänsch, T. W. & Hemmerich, A. Local and global properties of light-bound atomic lattices investigated by Bragg diffraction. *Phys. Rev. A* **58,** 4647–4661 (1998).
25. Miyake, H. *et al.* Bragg scattering as a probe of atomic wave functions and quantum phase transitions in optical lattices. *Phys. Rev. Lett.* **107,** 175302 (2011).
26. Corcovilos, T. A., Baur, S. K., Hitchcock, J. M., Mueller, E. J. & Hulet, R. G. Detecting antiferromagnetism of atoms in an optical lattice via optical Bragg scattering. *Phys. Rev. A* **81,** 013415 (2010).
27. Werner, F., Parcollet, O., Georges, A. & Hassan, S. R. Interaction-induced adiabatic cooling and antiferromagnetism of cold fermions in optical lattices. *Phys. Rev. Lett.* **95,** 056401 (2005).
28. Fuchs, S. *et al.* Thermodynamics of the 3D Hubbard model on approaching the Néel transition. *Phys. Rev. Lett.* **106,** 030401 (2011).
29. Mathy, C. J. M. & Huse, D. A. Accessing the Néel phase of ultracold fermionic atoms in a simple-cubic optical lattice. *Phys. Rev. A* **79,** 063412 (2009).
30. Köhl, M. Thermometry of fermionic atoms in an optical lattice. *Phys. Rev. A* **73,** 031601 (2006).

Acknowledgements This work was supported under ARO grant no. W911NF-13-1-0018 with funds from the DARPA OLE programme, NSF, ONR, the Welch Foundation (grant no. C-1133), and an ARO-MURI grant no. W911NF-14-1-003. T.P. acknowledges support from CNPq, FAPERJ, and the INCT on Quantum Information. R.T.S. acknowledges support from the Office of the President of the University of California.

Author Contributions The experimental work was performed by R.A.H., P.M.D., T.-L.Y., X.L. and R.G.H., while T.P., E.K., P.M.D., R.T.S., N.T. and D.A.H. performed the theory needed to extract temperatures from the data and provided overall theoretical guidance. All authors contributed to the writing of the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to R.G.H. (randy@rice.edu).

METHODS

Preparation. ^{6}Li atoms are first captured and cooled in a magneto-optical trap (MOT) operating at 671 nm. They are further cooled in a second MOT stage employing 323 nm light near resonant with the $2S \rightarrow 3P$ transition. As described previously[20], these atoms are laser cooled into a large-volume optical dipole trap (ODT) where a balanced spin mixture of the states $|\uparrow\rangle = |2S_{1/2}; F = 1/2, m_F = +1/2\rangle$ and $|\downarrow\rangle = |2S_{1/2}; F = 1/2, m_F = -1/2\rangle$ is produced.

Once the large-volume ODT is loaded, we set the magnetic field to 340 G ($a_s \approx -289a_0$) to perform evaporative cooling. The intensities of the lattice beams (1,064 nm) in dimple configuration (with the polarization of each retroreflection perpendicular to that of each input beam) are turned on in 1 s. The depth of the dimple, which at this point is only a small perturbation on the ODT, is adjusted to control the final atom number in the experiment. The depth of the ODT is then ramped to zero in 5.5 s to evaporatively cool the atoms into the dimple. To produce a final sample with repulsive interactions, the magnetic field is increased to 595 G ($a_s \approx +326a_0$) in a 5 ms linear ramp starting 3 s into the evaporation trajectory. Owing to the small volume of the dimple relative to the ODT, evaporation into the dimple is efficient and deeply degenerate samples are reliably produced.

We measure T/T_F in the dimple trap by fitting the density profile, after 0.5 ms of time-of-flight, to a Thomas–Fermi distribution[31]. The magnetic field is tuned to 528 G to make the gas non-interacting before the measurement. For the experiments reported here, the final dimple depths are in the range between $0.325E_r$ and $0.5E_r$ per axis, resulting in N in the range $(1.0\text{–}2.5) \times 10^5$. The measured value $T/T_F = 0.04 \pm 0.02$ is independent of N within this range. The uncertainty in T/T_F is the standard deviation of the fitted value for at least six independent realizations. Here as elsewhere no statistical methods were used to predetermine sample size.

Compensated optical lattice. The experiment takes place in a compensated simple cubic optical lattice potential that can be expressed as

$$V_{3D}(x,y,z) = V_{1D}(x;y,z) + V_{1D}(y;z,x) + V_{1D}(z;x,y)$$

where

$$V_{1D}(x;y,z) = V_L(x;y,z) + V_C(x;y,z)$$

and V_L, V_C are the potentials produced by the lattice (1,064 nm) and compensation (532 nm) beams, respectively:

$$V_L(x;y,z) = -v_0 \exp\left[-2\frac{y^2+z^2}{w_L^2}\right]\cos^2\left(\frac{2\pi}{\lambda}x\right)$$

$$V_C(x;y,z) = g_0 \exp\left[-2\frac{y^2+z^2}{w_C^2}\right].$$

Here, v_0 is the lattice depth and g_0 is the compensation ($v_0, g_0 > 0$). A schematic of the compensated lattice, and the spatial variation of the Hubbard parameters due to the finite lattice beam waists, are shown in Extended Data Fig. 1.

The beam waists ($1/e^2$ radius) of the three axes are calibrated independently by phase modulation spectroscopy of each lattice beam and by measuring the frequency of breathing mode oscillations. The waists are found to be (up to a $\pm 5\%$ systematic uncertainty) $w_L = (47, 47, 44)$ μm and $w_C = (42, 41, 40)$ μm, for beams propagating along x, y, z, respectively.

Lattice loading. To load the lattice from the dimple trap, we first rotate the polarization of the retroreflected beams parallel to that of the input beams in 100 ms. In the following 25 ms, we increase the lattice depth to $2.5E_r$ and ramp the magnetic field to set the final value of U_0/t_0. The lattice depth is then ramped to $7.0E_r$ in 15 ms.

Throughout the process of loading the lattice from the dimple, the power of the compensating beams is adjusted in order to maintain the peak density of the sample at $n \approx 1$. At the final lattice depth of $v_0 = 7.0E_r$, the average compensation per beam is $g_0 = 3.7E_r$. The value of g_0 for each beam is adjusted slightly from this average in order to create samples that appear spherically symmetric.

Round-trip T/T_F measurements. After loading the atoms into the $7E_r$ lattice we wait for a hold time t_h and then reverse the lattice loading ramps to return to the harmonic dimple trap and measure T/T_F. This measurement, shown in Extended Data Fig. 3, sets an upper limit on the entropy of the system in the lattice, and is also a measure of the heating rate of the system in the lattice.

Temperature dependence of S_π. In Extended Data Fig. 4 we show S_π as a function of hold time in the lattice t_h and observe that it decays for longer hold times, as expected from the increase in T/T_F. Although the preparation of the sample and the final potential are somewhat different for the data presented in Extended Data Figs 3 and 4, the data support the contention that the Bragg signal decreases with increasing T.

Variation of N to maximize S_π. The global chemical potential μ_0 must be increased for larger U_0/t_0 to guarantee the formation of a Mott plateau in the trap. A larger μ_0 results in larger atom number. N is adjusted to maximize the Bragg signal for each experimental value of U_0/t_0 in Fig. 4. We adjust N by tuning the depth of the dimple trap in which degeneracy is achieved before loading the atoms into the lattice. The optimal value of N as a function of U_0/t_0 is shown in Extended Data Fig. 2.

Spin structure factor measurement. We measure the spin structure factor at two different values of the momentum transfer $\boldsymbol{Q}$ given by

$$\boldsymbol{\pi} = \frac{2\pi}{a}(-0.5, -0.5, +0.5)$$

$$\boldsymbol{\theta} = \frac{2\pi}{a}(+0.396, -0.105, -0.041),$$

where $a = \lambda/2$ is the lattice spacing.

We detect the scattered light using two separate cameras as the cloud is illuminated with the Bragg probe beam for 1.7 μs. The Bragg probe beam is a collimated Gaussian beam with a waist of 450 μm and 250 μW of power, resulting in an intensity $I_p = 79\ \mathrm{mW\ cm^{-2}}$. The intensity of the probe determines the on-resonance saturation parameter $s_0 = I_p\left|\hat{\boldsymbol{e}}_p \cdot \hat{\boldsymbol{e}}_{-1}\right|^2 \Big/ \left(\frac{\pi hc\Gamma}{2\lambda_0^3}\right) = 15.5$, where c is the speed of light, $\hat{e}_p$ is the polarization of the probe light, $\hat{e}_{-1}$ is the unit vector in the direction of the dipole matrix element of the transition, $\lambda_0 = 671$ nm is the wavelength of the transition, and Γ is its linewidth. The polarization of the incident light in our experiment is linear and perpendicular to the quantization axis, so $\left|\hat{\boldsymbol{e}}_p \cdot \hat{\boldsymbol{e}}_{-1}\right|^2 = 1/2$. The Bragg probe detuning is set between the two spin states, such that $\Delta = |\Delta_\uparrow| = |\Delta_\downarrow| = 6.4\Gamma$, where $\Delta_\uparrow$ and $\Delta_\downarrow$ are the detunings from the two spin states.

The spin structure factor is defined in equation (1) as a sum over lattice sites i, j. By quickly ramping the lattice depth to $v_0 = 20E_r$, the state of the system is projected into a product state, where the wavefunction of each atom is localized at a lattice site. Hence, we can write $S_{\boldsymbol{Q}}$ as a sum over particles m, n:

$$S_{\boldsymbol{Q}} = \frac{4}{N}\sum_{m,n} e^{i\boldsymbol{Q}\cdot(\boldsymbol{R}_m - \boldsymbol{R}_n)}\langle\sigma_z\rangle_m\langle\sigma_z\rangle_n$$

where $\langle\sigma_z\rangle_n$ is the z component of the spin of the nth atom.

When illuminated with the probe light, each atom can be considered as an independent scatterer, and the intensity at the detector can be obtained by summing the field contributions from the individual atoms and squaring the total field. We assume that the spatial wavefunction of all atoms is the harmonic oscillator ground state in a lattice site of depth v_0, and that it does not change during the measurement. The resulting intensity at the detector is given by

$$I_{\boldsymbol{Q}}(\tau) = \frac{As_0/2}{4\delta^2 + s_0}N + e^{-2W_{\boldsymbol{Q}}(\tau)}\frac{2As_0\delta^2}{\left(4\delta^2+s_0\right)^2}\sum_{\substack{m,n \\ m\neq n}} 4\langle\sigma_z\rangle_m\langle\sigma_z\rangle_n e^{i\boldsymbol{Q}\cdot(\boldsymbol{R}_n - \boldsymbol{R}_m)} \quad (2)$$

where $\delta = \Delta/\Gamma$, and $A = \frac{3}{8\pi}\frac{\hbar ck\Gamma}{r_D^2}|\boldsymbol{\Lambda}|^2$. Here $\boldsymbol{\Lambda}$ is the polarization vector of the scattered field, $\boldsymbol{\Lambda} = \hat{\boldsymbol{n}} \times (\hat{\boldsymbol{n}} \times \hat{\boldsymbol{e}}_{-1})$, where $\hat{\boldsymbol{n}}$ is a unit vector pointing in the direction of the detector, which is located at a distance r_D from the sample.

In equation (2) the first term arises from uncorrelated scattering by the atoms, while the second term represents the interference due to magnetic correlations. We can identify the spin structure factor in the interference term as

$$\sum_{\substack{m,n \\ m\neq n}} 4\langle\sigma_z\rangle_m\langle\sigma_z\rangle_n e^{i\boldsymbol{Q}\cdot(\boldsymbol{R}_n - \boldsymbol{R}_m)} = N\left(S_{\boldsymbol{Q}} - 1\right)$$

and obtain

$$S_{\boldsymbol{Q}} = 1 + C_{\boldsymbol{Q}}(\tau)\left(\frac{I_{\boldsymbol{Q}}(\tau)}{I_{\boldsymbol{Q}\infty}} - 1\right)$$

where $I_{\boldsymbol{Q}\infty} = \frac{As_0/2}{4\delta^2+s_0}N$, and the correction factor is $C_{\boldsymbol{Q}}(\tau) = e^{2W_{\boldsymbol{Q}}(\tau)}\left(1 + \frac{s_0}{4\delta^2}\right)$. In the experiment we obtain $S_{\boldsymbol{Q}}$ by combining measurements of the scattered intensity *in situ* ($\tau = 0$) and after sufficiently long time-of-flight ($\tau = 6$ μs). The correction factor takes the values $C_{\boldsymbol{\pi}}(\tau = 0) = 1.52$ for $\boldsymbol{Q} = \boldsymbol{\pi}$ and $C_{\boldsymbol{\theta}}(\tau = 0) = 1.18$ for $\boldsymbol{Q} = \boldsymbol{\theta}$.

Time-of-flight. After the atoms are released in time-of-flight, the Debye–Waller factor decays as the atomic wavefunctions expand, resulting in a corresponding decay of the Bragg scattered intensity. For a lattice of depth v_0

$$e^{-2W_{\boldsymbol{Q}}(\tau)} = e^{-2W_{\boldsymbol{Q}}(\tau=0)}\exp\left[-\frac{\sqrt{v_0/E_r}}{2}\left(\frac{|\boldsymbol{Q}|h}{2ma}\right)^2\tau^2\right].$$

This equation was used to calculate the solid grey line in Fig. 2. The average value of the Debye–Waller factor during the duration of the Bragg exposure

$$(1.7\mu s)^{-1}\int_{\tau}^{\tau+1.7\mu s} e^{-2W_{\boldsymbol{Q}}(\tau')}d\tau'$$

is used to calculate the dashed grey line in Fig. 2.

The data shown in Fig. 2 was taken at $U_0/t_0 = 13.4$ with $N = 2.5\times10^5$ atoms. This value of N is above the optimal value, so the ratio of $I_{\pi 0}/I_{\pi\infty}$ in Fig. 2 gives $S_{\boldsymbol{\pi}} \approx 1.4$, which is less than the expected optimal value of $S_{\boldsymbol{\pi}}$ from Fig. 4.

Momentum transferred from the probe to the atoms. As mentioned above, we assume that the spatial wavefunction of the atoms remains unchanged for the duration of the exposure. For this assumption to be valid, the Lamb–Dicke parameter $\eta^2 = \dfrac{h^2/(2m\lambda_0^2)}{2E_r\sqrt{v_0/E_r}}$ needs to be $\ll 1$. In the $20E_r$ lattice, $\eta^2 = 0.27$, meaning that approximately one out of every four photons scattered will excite an atom to the second band of the lattice. An atom in the second band has larger position uncertainty and hence a smaller Debye–Waller factor, which reduces its contribution to the Bragg scattering signal.

The total number of photons scattered per atom is given by $N_p = t_{exp}\Gamma\dfrac{s_0/2}{s_0+4\delta^2}$, where the duration of the probe is $t_{exp} = 1.7\,\mu s$. For $s_0 = 15.5$ and $\delta = 6.4$, $N_p = 2.7$, thus justifying the assumption that the atoms remain in the lowest band during the pulse.

For the Bragg scattering measurements performed after time-of-flight, the momentum transferred from the probe to the atoms plays a more important role, since the atoms are not trapped and will recoil after every photon scatter. Despite this, we still see a good agreement between the observed decay of the Bragg scattering signal and the decay expected for a Heisenberg-limited wave packet, as shown in Fig. 1. We have also performed non-spin-sensitive Bragg scattering measurements from the 010 planes of the lattice and observe the same agreement, justifying that momentum transfer from the probe to the atoms can be neglected for the exposure times used.

Optical density. A low optical density of the sample is important so that the probe is unattenuated through the atom cloud, and multiple scattering events of the Bragg scattered photons are limited[26]. The optical density can be approximated as

$$\mathrm{OD}\approx\frac{\sigma_0\left|\hat{\boldsymbol{e}}_p\cdot\hat{\boldsymbol{e}}_{-1}\right|^2}{4\delta^2+s_0}\frac{1}{a^2}\left(\frac{3N}{4\pi}\right)^{1/3}$$

where $\sigma_0 = 3\lambda_0^2/2\pi$. With $s_0 = 15.5$, $\delta = 6.4$ and $N = 1.8\times10^5$ atoms we have $\mathrm{OD}\approx 0.072$. At this value we do not expect significant corrections to the spin structure factor measurement due to the attenuation of the probe. We have not included any corrections in our measurement due to finite optical density effects.

Light collection. We collect Bragg scattered light in the $\boldsymbol{\pi}$ direction over a full angular width of 110 mrad, given by a 2.5 cm diameter collection lens located 23 cm away from the atoms. In the $\boldsymbol{\theta}$ direction, light is collected by a 2.5 cm diameter lens placed 8 cm away from the atoms, corresponding to a full angular width of 318 mrad. The scattered light in each of the directions is focused to a few pixels on the cameras, so no additional angular information is obtained. For $N = 1.8\times10^5$, $s_0 = 15.5$, $\Delta = 6.4\,\Gamma$ and a 1.7 μs pulse, the detector in the $\boldsymbol{\pi}$ direction collects approximately 1,300 photons, whereas the detector in the $\boldsymbol{\theta}$ direction collects approximately 10^4 photons. The noise floor from readout, dark current and background light per shot has a variance equivalent to approximately 250 photons in the $\boldsymbol{\pi}$ direction and 1,000 photons in the $\boldsymbol{\theta}$ direction.

Data averaging. The signals we detect are small enough that an uncorrelated sample may, in a single shot, produce a scattering signal as large as the ones produced by samples with AFM correlations. To obtain a reliable measurement of $S_{\boldsymbol{\pi}}$ we average at least 40 *in situ* shots to obtain $I_{\boldsymbol{Q}0}$ and at least 40 time-of-flight shots to obtain $I_{\boldsymbol{Q}\infty}$.

We estimate the expected variance on $S_{\boldsymbol{\pi}}$ by considering a randomly ordered sample in which $e^{i\boldsymbol{\pi}\cdot\boldsymbol{R}_n}2\langle\sigma_z\rangle_n$ is equal to $+1$ or -1 with equal probability. $S_{\boldsymbol{\pi}}$ can be written as

$$S_{\boldsymbol{\pi}} = \left|\sum_n e^{i\boldsymbol{\pi}\cdot\boldsymbol{R}_n}\frac{2\langle\sigma_z\rangle_n}{\sqrt{N}}\right|^2$$

which is equivalent to the square of the distance travelled on an unbiased random walk with step size $1/\sqrt{N}$. The mean and standard deviation can then be readily calculated: $\overline{S_{\boldsymbol{\pi}}} = 1$ and $\sqrt{\mathrm{Var}(S_{\boldsymbol{\pi}})} = \sqrt{2}$, where $\mathrm{Var}(S_{\boldsymbol{\pi}})$ denotes the variance of the random variable $S_{\boldsymbol{\pi}}$. With a standard deviation that is larger than the mean value, a considerable number of shots needs to be taken in order to obtain an acceptable error in the mean. The standard error of the mean for 40 shots will be $\sqrt{2/40} = 0.22$, consistent with what we obtain in the experiment (see Fig. 4).

Numerical calculations. DQMC and numerical linked-cluster expansion (NLCE) calculations are used to obtain the local values of the thermodynamic quantities in our trap, including the density, entropy, and the spin structure factor. DQMC calculations for arbitrary chemical potential (and hence density) can be obtained reliably down to temperatures slightly above the Néel temperature for a given $U/t \lesssim 9$. For stronger interactions, intermediate values of n become inaccessible to DQMC owing to the sign problem, in which case we rely on the NLCE to obtain values of the thermodynamic quantities for arbitrary chemical potential down to temperatures as low as $T/t = 0.40$.

DQMC results for a $6\times6\times6$ lattice were obtained with the methodology described in refs 7 and 32. Inverse temperature discretization $\Delta\tau = \beta/L$, where $\beta = 1/T$ and $L = 20\beta t$, is sufficiently small that Trotter corrections are substantially less than statistical error bars. DQMC data were obtained with 1,000 sweeps through the lattice for equilibration, and between 5,000 (small U and high T) and 200,000 (large U and low T) sweeps for measurements. Finite-size effects were assessed by comparing DQMC results for $6\times6\times6$ and $8\times8\times8$ lattices. Differences are only appreciable when the spin structure factor per lattice site, $s_{\boldsymbol{\pi}} > 5$. The local value of $s_{\boldsymbol{\pi}}$ is always less than 4 in calculations shown here, so DQMC results in a $6\times6\times6$ lattice are sufficient for the comparison with theory.

In NLCEs, an extensive property of the lattice model per site in the thermodynamic limit is expressed in terms of contributions from finite clusters that can be embedded in the lattice. NLCEs use the same basis as high-temperature expansions, however, properties of clusters are calculated via exact diagonalization, as opposed to a perturbative expansion in powers of the inverse temperature[8,33]. The site-based NLCE for the Hubbard model[34] is implemented here for a 3D lattice and carried out to the eighth order for all thermodynamic quantities, except for $S_{\boldsymbol{\theta}}$, where due to the reduced symmetry, only seven orders were obtained. Within its region of convergence ($T/t \gtrsim 1.5$ for any n and U), NLCE results do not contain any systematic or statistical errors. The convergence region extends to much lower T/t at $n = 1$ and generally improves by increasing the interaction strength. At lower T/t, we take advantage of numerical resummations, such as Euler and Wynn transformations[33], to obtain an estimate. The NLCE provides a fast tool, which, given the value of U/t, generates results on a dense temperature and chemical potential grid in a single run.

Local density approximation. The local density approximation, which has been previously shown to agree well with *ab initio* DQMC simulations of the trapped Hubbard Hamiltonian[35], was used to calculate the trap profiles of the different thermodynamic quantities. The spin structure factor $S_{\boldsymbol{Q}}$ is obtained from the trap profile of the spin structure factor per lattice site as

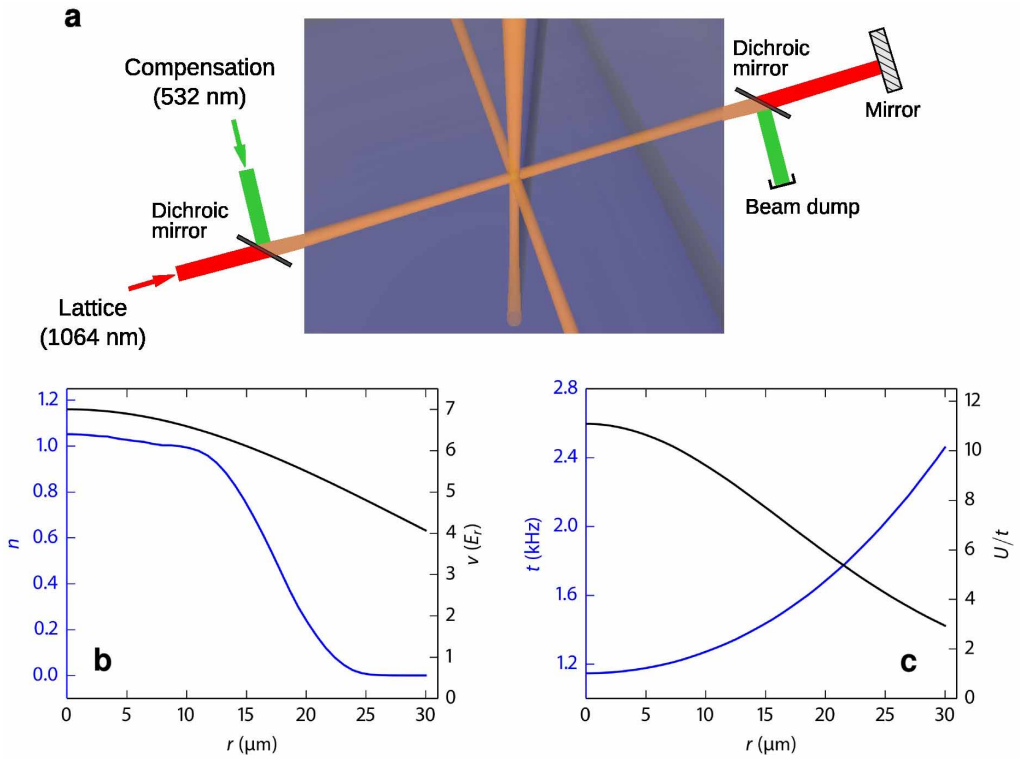
$$S_{\boldsymbol{Q}} = \frac{1}{Na^3}\int s_{\boldsymbol{\pi}}d^3r$$

For the numerical calculations we set T and μ_0; local values of U/t, T/t, and the local chemical potential μ/t are calculated using the known trap potential. The local values of the thermodynamic quantities are then obtained by interpolation from NLCE and DQMC results for a homogeneous system calculated in a (U/t, T/t, μ/t) grid. Radial profiles for the local value of U/t, T/t, and μ/t along a body diagonal of the lattice were used and spherical symmetry assumed.

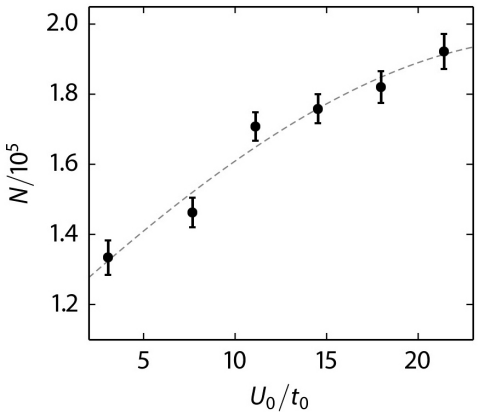
Entropy. In Fig. 4 we compare the experimental results at various U_0/t_0 with calculations at constant T. Since ultracold atoms are isolated systems, a constant value of the overall entropy per particle $S/(Nk_B)$ may be more appropriate. We find that over the range $10 < U_0/t_0 < 15$, where AFM correlations are largest, $S/(Nk_B)$ does not vary significantly with U_0/t_0, at constant T (Extended Data Fig. 6). This implies that we do not expect adiabatic cooling for stronger interactions[18,27], and thus the curves at constant T are suitable to describe the experimental data.

Code availability. The codes used for DQMC and NLCE calculations are available by request from the authors.

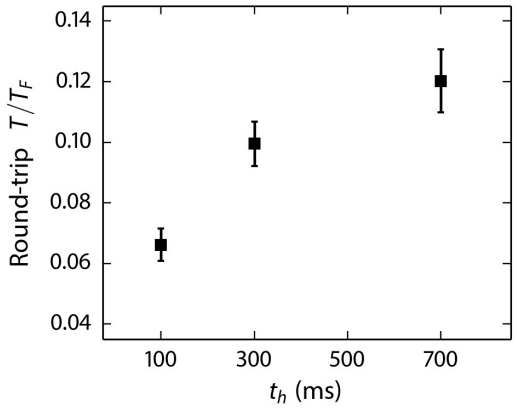
31. Butts, D. A. & Rokhsar, D. S. Trapped Fermi gases. *Phys. Rev. A* **55,** 4346–4350 (1997).
32. Paiva, T., Scalettar, R., Randeria, M. & Trivedi, N. Fermions in 2D optical lattices: temperature and entropy scales for observing antiferromagnetism and superfluidity. *Phys. Rev. Lett.* **104,** 066406 (2010).
33. Tang, B., Khatami, E. & Rigol, M. A short introduction to numerical linked-cluster expansions. *Comput. Phys. Commun.* **184,** 557–564 (2013).
34. Khatami, E. & Rigol, M. Thermodynamics of strongly interacting fermions in two-dimensional optical lattices. *Phys. Rev. A* **84,** 053611 (2011).
35. Chiesa, S., Varney, C. N., Rigol, M. & Scalettar, R. T. Magnetism and pairing of two-dimensional trapped fermions. *Phys. Rev. Lett.* **106,** 035301 (2011).



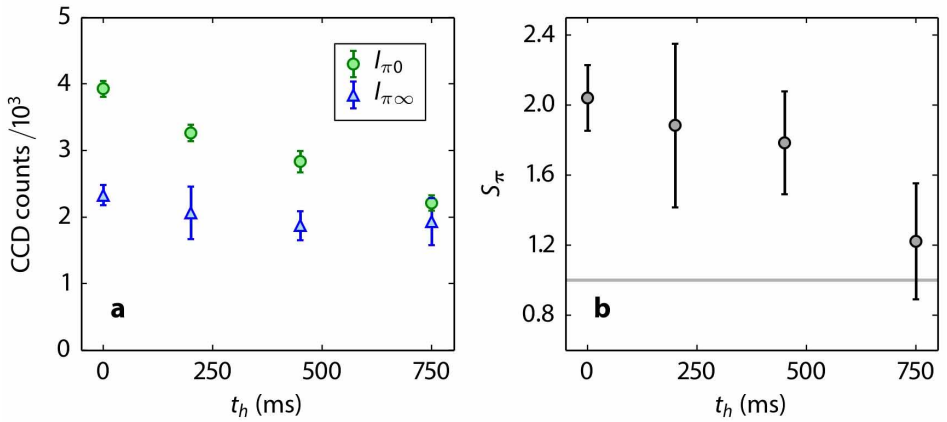
Extended Data Figure 1 | Compensated optical lattice. **a**, Schematic of the compensated optical lattice set-up. Along each axis, the radial confinement of the lattice is compensated with a repulsive compensation beam which is combined with the lattice beam using a dichroic mirror. The compensation beam co-propagates with the lattice beam but is not retroreflected; instead a dichroic mirror before the retro-reflection mirror is used to direct the compensation beam to a beam dump. **b**, The local value of the lattice depth v (black line; right-hand y axis) is shown as a function of distance from the centre along a body diagonal of the lattice. Owing to the finite extent of the lattice beams, v varies across the density profile of the cloud. The density n, calculated for $U_0/t_0 = 11.1$ at $T/t_0 = 0.60$, is shown (blue line; left-hand y axis). **c**, The inhomogeneity in v results in spatially varying Hubbard parameters t (blue line; left-hand y axis) and U/t (black line; right-hand y axis).



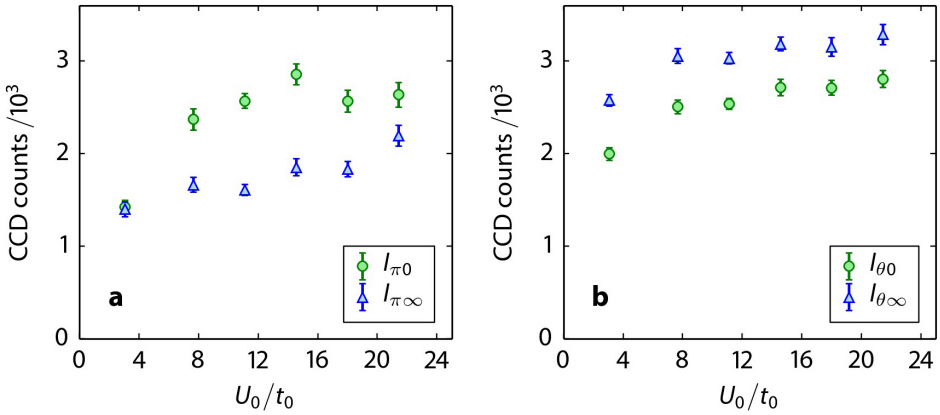
Extended Data Figure 2 | Atom number for the data in Fig. 4. Atom number N which maximizes S_π as a function of U_0/t_0. We control N by adjusting the depth of the dimple trap. Using a linear calibration between the depth of the dimple trap and the final atom number, we obtain the value of N corresponding to the data in Fig. 4. The error bars correspond to the s.e.m. of the dimple depths used in at least 40 *in situ* and 40 time-of-flight realizations of the experiment, corresponding to the data in Fig. 4. The line is a third-order polynomial fit, which is used to interpolate the value of N for numerical calculations shown in Fig. 4.



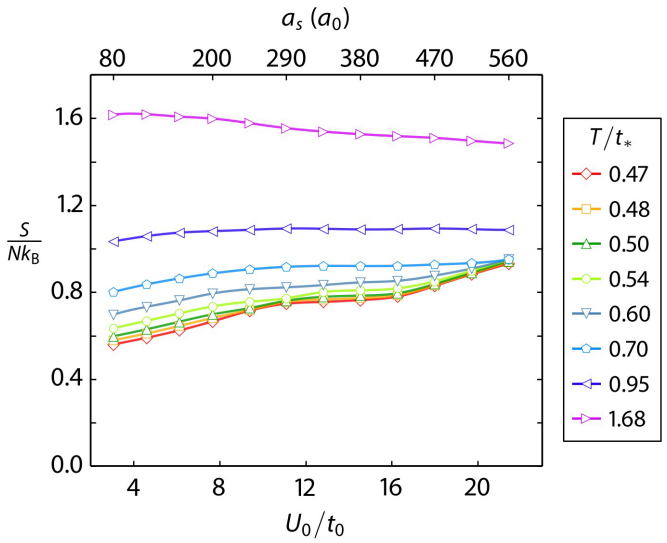
Extended Data Figure 3 | Round-trip temperature measurements. Measurement of the round-trip T/T_F versus hold time t_h in a compensated lattice with $v_0 = 7E_r$ and $g_0 = 3.7E_r$. The duration of the loading ramps is not included in t_h. The scattering length is $326a_0$, which corresponds to $U_0/t_0 = 12.5$. Error bars are the s.e.m. of six independent realizations. The temperature in the dimple trap before loading into the lattice is $T/T_F = 0.04 \pm 0.02$.



Extended Data Figure 4 | Bragg signal decay with hold time. **a**, Detected counts (from CCD camera) versus t_h, measured for momentum transfer $\boldsymbol{Q} = \boldsymbol{\pi}$ for an *in situ* sample ($I_{\pi 0}$, green circles) and after decay of the Debye–Waller factor ($I_{\pi\infty}$, blue triangles). For longer hold times, the Bragg-scattered intensity $I_{\pi 0}$ decays to match $I_{\pi\infty}$, reflecting the absence of AFM correlations in a sample at higher T. **b**, The spin structure factor S_π corresponding to the scattered intensities shown in **a**. For these measurements the scattering length is $200a_0$, corresponding to $U_0/t_0 = 7.7$ in a $7E_r$ deep lattice. The compensation is $g_0 = 4.05E_r$, different from that used for the data in Fig. 4. The increased compensation requires a larger atom number to realize an $n \approx 1$ shell in the cloud. The atom number used here is 2.6×10^5 atoms. The duration of the Bragg probe is 2.7 μs for these data. Error bars in **a** are the s.e.m. of at least 5 measurements for $I_{\pi\infty}$ and at least 10 measurements for $I_{\pi 0}$. Error bars in **b** are obtained from the s.e.m. of the measured intensities and equation (2).



Extended Data Figure 5 | Detected counts for measurement of spin structure factor in Fig. 4. a, Detected counts versus U_0/t_0, measured for momentum transfer $\boldsymbol{Q} = \boldsymbol{\pi}$ for an *in situ* sample ($I_{\pi 0}$, green circles), and after decay of the Debye–Waller factor ($I_{\pi\infty}$, blue triangles). As U_0/t_0 increases we use a larger atom number to optimize the Bragg signal. $I_{\pi\infty}$ and $I_{\pi 0}$ both increase with U_0/t_0 owing to the larger N, but $I_{\pi 0}$ shows an additional enhancement due to the presence of AFM correlations. **b**, Detected counts versus U_0/t_0, measured for momentum transfer $\boldsymbol{Q} = \boldsymbol{\theta}$ for an *in situ* sample ($I_{\theta 0}$, green circles), and after decay of the Debye–Waller factor ($I_{\theta\infty}$, blue triangles). For $\boldsymbol{Q} = \boldsymbol{\theta}$ most of the dependence for both the *in situ* and time-of-flight intensities is due to the changing N. Error bars in both **a** and **b** are the s.e.m. of at least 40 measurements. The overall count rate is higher for $\boldsymbol{Q} = \boldsymbol{\theta}$ owing to the different collection efficiency and gain settings of the CCD camera.



Extended Data Figure 6 | Entropy per particle at constant *T*. Overall entropy per particle $S/(Nk_B)$ as a function of U_0/t_0 for the calculations at various T/t_* shown in Fig. 4 (lines are guides to the eye). For the lowest temperatures, $S/(Nk_B)$ does not vary significantly over the range of U_0/t_0 covered by the experiment, justifying the treatment at constant T. A value of $S/(Nk_B) \approx 0.76$ is obtained for the temperature determined from the data in Fig. 4.

LETTER

doi:10.1038/nature14172

Decrease in CO_2 efflux from northern hardwater lakes with increasing atmospheric warming

Kerri Finlay[1], Richard J. Vogt[1]†, Matthew J. Bogard[1]†, Björn Wissel[2], Benjamin M. Tutolo[3], Gavin L. Simpson[2] & Peter R. Leavitt[1,2]

Boreal lakes are biogeochemical hotspots that alter carbon fluxes by sequestering particulate organic carbon in sediments[1,2] and by oxidizing terrestrial dissolved organic matter to carbon dioxide (CO_2) or methane through microbial processes[3,4]. At present, such dilute lakes release ~1.4 petagrams of carbon annually to the atmosphere[3,4], and this carbon efflux may increase in the future in response to elevated temperatures[5] and increased hydrological delivery of mineralizable dissolved organic matter to lakes[6,7]. Much less is known about the potential effects of climate changes on carbon fluxes from carbonate-rich hardwater and saline lakes that account for about 20 per cent of inland water surface area[4,8]. Here we show that atmospheric warming may reduce CO_2 emissions from hardwater lakes. We analyse decadal records of meteorological variability, CO_2 fluxes and water chemistry to investigate the processes affecting variations in pH and carbon exchange[9,10] in hydrologically diverse lakes of central North America. We find that the lakes have shifted progressively from being substantial CO_2 sources in the mid-1990s to sequestering CO_2 by 2010, with a steady increase in annual mean pH. We attribute the observed changes in pH and CO_2 uptake to an atmospheric-warming-induced decline in ice cover in spring that decreases CO_2 accumulation under ice, increases spring and summer pH, and enhances the chemical uptake of CO_2 in hardwater lakes. Our study suggests that rising temperatures do not invariably increase CO_2 emissions from aquatic ecosystems.

Boreal lakes are important in global carbon (C) cycles because they receive ~2.9 Pg C per year from terrestrial sources[3,4], permanently bury ~0.6 Pg per year as particulate C (refs 1, 2), and mineralize up to 50% of the remainder to CO_2 and methane[4] through bacterial activity in the water column[11] and sediments[12]. In general, dilute unproductive lakes release more gaseous C than is fixed by aquatic photosynthesis[11,13,14], whereas net CO_2 uptake occurs in some productive basins when elevated nutrient influx intensifies primary production and labile organic C is incompletely mineralized by bacteria[3,4,15]. At present, the magnitude of C fluxes from boreal lakes is similar to those arising from global deforestation, oceanic CO_2 sequestration and net terrestrial production[4]; however, future mineralization of organic matter is predicted to intensify under a warmer[5] or wetter climate[6,7].

Less is known about how solute-rich hardwater lakes influence planetary C fluxes[4,8], despite accounting for ~50% of inland waters by volume[16] (23% by area)[8], in part because pH regulates inter-annual variation in atmospheric CO_2 exchange at these sites independently of microbial metabolism during summer[8,9], and because controls of inter-annual variation in pH are poorly understood[9,10]. Typically, hardwater lakes are alkaline ($8 < pH < 11$), rich in dissolved inorganic C (DIC) derived from catchment sources of HCO_3^- and CO_3^{2-}, and evade (release) much more CO_2 (up to 200 mmol C m^{-2} d^{-1}) than do boreal lakes (up to 60 mmol m^{-2} d^{-1}) when pH values are below 9.0 (refs 4, 8–10). At higher pH, CO_2 is converted to HCO_3^- and CO_3^{2-} (ref. 17), partial pressure of CO_2 (p_{CO_2}) declines to below atmospheric values, and hardwater lakes sequester atmospheric CO_2 (refs 8–10). Furthermore, DIC-rich hardwater and saline lakes exhibit a high degree of spatial synchrony in mean summer pH[18,19] and can rapidly vary the direction and magnitude of CO_2 flux[9]. Thus, a better understanding of the mechanisms regulating inter-annual variation in pH and carbon processing of hardwater lakes is essential to quantify the contribution of these ecosystems to the global carbon cycle[4,20].

We analysed 16 years of meteorological and limnological data collected every two weeks during May to August from six lakes, a 28-year record of weekly chemical determinations at one lake, and surveys of water chemistry in an additional 20 (seasonal) to 70 (annual) lakes to identify factors regulating inter-annual variation in lake pH and CO_2 flux within a 236,000 km^2 region of the Northern Great Plains of North America (Extended Data Fig. 1). Our grassland study region represents more than 40% of all cultivated land in Canada and is composed mainly (75%) of agricultural fields and pastures, particularly within the 52,000 km^2 Qu'Appelle River catchment (50° 00′–51° 30′ N, 101° 30′–107° 10′ W) of southern Saskatchewan[21]. Study lakes within this drainage basin vary tenfold in most morphometric, hydrological and limnological features (Extended Data Table 1), include both reservoirs (Wascana and Diefenbaker lakes) and sites with limited hydrological outflow (Last Mountain Lake), yet are all alkaline (mean summer pH ~8.8; 30–60 mg DIC l^{-1}) and well mixed (except occasionally stratified Katepwa Lake) and have common plankton composition and trophic relationships[10,21].

Analysis of 16 years of water chemistry and C flux data revealed that Qu'Appelle lakes have shifted progressively from being large CO_2 sources in the mid-1990s to sequestering substantial amounts of CO_2 at present (Fig. 1). The annual pH of these lakes has steadily increased from 8.3 ± 0.1 in 1995 to 9.2 ± 0.1 in 2010 (means ± s.e.m.; $n = 6$; Fig. 1a), whereas total inorganic carbon (TIC) (Fig. 1b), hydrological influx[9] (not shown) and lake production[9] (not shown) have remained essentially unchanged. The consequence of these shifts is that aquatic p_{CO_2} has declined nearly tenfold in all lakes (Fig. 1c), atmospheric CO_2 evasion has been decreased by nearly 100 g C m^{-2} per summer (Fig. 1d), and lakes now sequester substantial quantities of CO_2 (37.4 ± 6.5 g C m^{-2} per summer) (Fig. 1d). Despite the marked physical, hydrological and chemical differences between lakes (Extended Data Table 1), inter-annual variation in pH and CO_2 parameters is highly coherent among ecosystems[9] and shows spatial patterns of synchrony that are characteristic of ecosystem regulation by energy influx (air temperature, irradiance) rather than by mass influx (precipitation, runoff, solutes)[19,22].

Principal components analysis suggests that the mean summer pH of Qu'Appelle lakes increased as a function of both spring and annual air temperatures, was correlated inversely with the duration and date of ice melt, and was uncorrelated with other measured meteorological variables (Extended Data Fig. 2a). In particular, pH was elevated under warmer atmospheric conditions, including those associated with a negative Southern Oscillation Index and positive (warm) phase of the Pacific Decadal Oscillation, both of which represent mild winters and reduced ice cover[23]. In contrast, mean summer pH was not correlated strongly with any measured aspect of lake chemistry, other than ammonium

[1]Limnology Laboratory, Department of Biology, University of Regina, Regina, Saskatchewan, Canada S4S 0A2. [2]Institute of Environmental Change and Society, University of Regina, Regina, Saskatchewan, Canada S4S 0A2. [3]Department of Earth Sciences, University of Minnesota, Minneapolis, Minnesota 55455, USA. †Present addresses: Department of Biology, Trent University, Peterborough, Ontario, Canada K9J 7B8 (R.J.V.); Département des Sciences Biologiques, Université du Québec à Montréal, Montréal, Québec, Canada H3C 3P8 (M.J.B.).

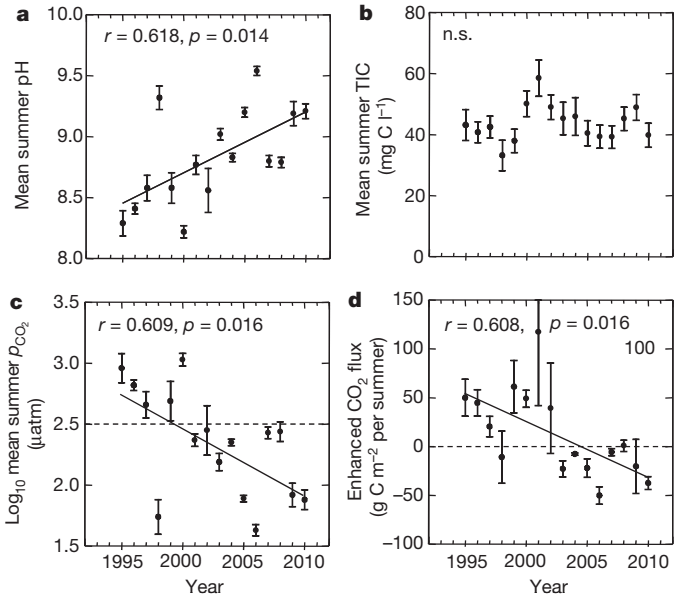


Figure 1 | Temporal changes in summer pH and CO_2 flux in six hardwater lakes of central Canada. **a**, Surface water pH; **b**, total inorganic carbon (TIC) concentration (mg C l^{-1}); **c**, $\log_{10}$ partial pressure of CO_2 (µatm); **d**, chemically enhanced flux of CO_2 (g C m^{-2} per summer). All time series are unweighted means and s.e.m. ($n = 6$). Least-squares regression analysis revealed linear increases in pH and declines in p_{CO_2} and CO_2 flux when conducted using years with complete summer sampling. Least-squares regression analyses exclude 2000, a year lacking samples during late July to September. Mean p_{CO_2} of the atmosphere (370 µatm) is indicated in **c** with a horizontal dashed line.

(NH_4^+) concentration (Extended Data Fig. 2b). Together these patterns are consistent with previous observations from other hardwater lakes and suggest that prolonged ice cover arising from cold winters favours increased CO_2 accumulation under ice and declines in under-ice pH after CO_2 hydration and carbonic acid formation[24,25].

Least-squares regression analysis of the decadal time series for Qu'Appelle lakes also showed that warmer winter temperatures were correlated negatively with both the duration of ice cover (Fig. 2a) and the date of ice melt (Fig. 2b), as has been noted elsewhere[23,26]. During years of prolonged cover, the date of ice melt is delayed by up to 20 days and the pH of Qu'Appelle lakes during spring (see below) and summer can be depressed by up to 1 pH unit (Fig. 2c). As shown in diverse lake districts, prolonged ice cover allows the accumulation of CO_2 from mineralized organic matter, which in turn hydrates to lower pH through formation of carbonic acid[27,28]. Furthermore, as pH is depressed, chemical equilibria dictate that a higher proportion of DIC is present as free CO_2, which can evade to the atmosphere[17]. In Qu'Appelle lakes, summer pH values below 9.0 were associated with substantial CO_2 evasion, whereas these lakes captured up to 50 g C m^{-2} during summers with a mean pH of >9.0 (Fig. 2d).

Detailed study of Buffalo Pound Lake, Saskatchewan, Canada, within the Qu'Appelle catchment illustrates the linkage between the duration of ice cover, the metabolic production of CO_2 and the depression of pH in spring and summer waters (Fig. 3). This lake has been monitored continuously at weekly intervals since 1979, with comprehensive chemical analysis from 1985 to 2003. Here we found a strong negative relationship between the duration of ice cover and the mean lake water pH during March, the month immediately preceding ice melt (Fig. 3a). In addition, spring pH was correlated positively with coeval determinations of oxygen content (Fig. 3b), suggesting that variations in mineralization of organic matter by microbes underlie both patterns[27,28]. Finally, we found a strong linear relationship between pH during March and values recorded in subsequent months (Fig. 3c), suggesting that variation in under-ice conditions can alter pH during the following summer.

Statistical and geochemical modelling of winter water chemistry from 1985 to 2003 reveals that metabolic production of CO_2 was the main control of inter-annual variation in the spring pH of Buffalo Pound Lake. For example, elastic net analysis explained 81% of observed deviance in winter pH and showed that microbial metabolism under ice was the principle predictor of the pH at spring ice melt, with a nearly fourfold greater standardized coefficient (0.14) than either HCO_3^- or Ca^{2+} (0.04), the only other significant and substantial model predictors (Extended Data Fig. 3). Similarly, geochemical modelling demonstrated that under-ice CO_2 evolution (O_2 decline × respiratory quotient of 1.2 = CO_2 production) was sufficient to depress the pH from values observed at ice formation (8.32 ± 0.06; mean ± s.e.m.) to those (7.83 ± 0.07) similar to values observed at the winter pH minimum (7.78 ± 0.08) or date of ice melt (8.06 ± 0.08). Finally, geochemical modelling revealed that spring ice melt resulted in a short-lived release of CO_2 but that the resultant increase in water-column pH was too small to uncouple the linear relationship between spring and summer pH (Extended Data Fig. 4). Together, these patterns suggest that variation in ice cover regulates the magnitude and direction of atmospheric CO_2 exchange by controlling spring and summer pH, altering the duration of the ice-free period and changing the proportion of time in which water-column pH is above or below the threshold of 9.0 (Fig. 2d).

Monitoring of other regional lakes since 2002 (refs 18, 29) has revealed that pronounced inter-annual variation in mean and seasonal pH is common and synchronous within the grassland region of central Canada (Extended Data Fig. 5). For example, the mean summer pH of Qu'Appelle lakes during 2002–2009 was highly correlated with that of ~20 DIC-rich closed-basin lakes within an independent 100,000 km^2 region (Extended Data Fig. 5a), whereas the rate of seasonal increase in pH was not significantly different between the two groups of lakes (Extended Data Fig. 5b). Furthermore, the chemical and hydrological properties of these closed-basin sites are representative of an additional ~50 DIC-rich hardwater and saline lakes surveyed during the past decade[18,29]. As shown elsewhere, inter-annual variation in pH within these closed-basin lakes

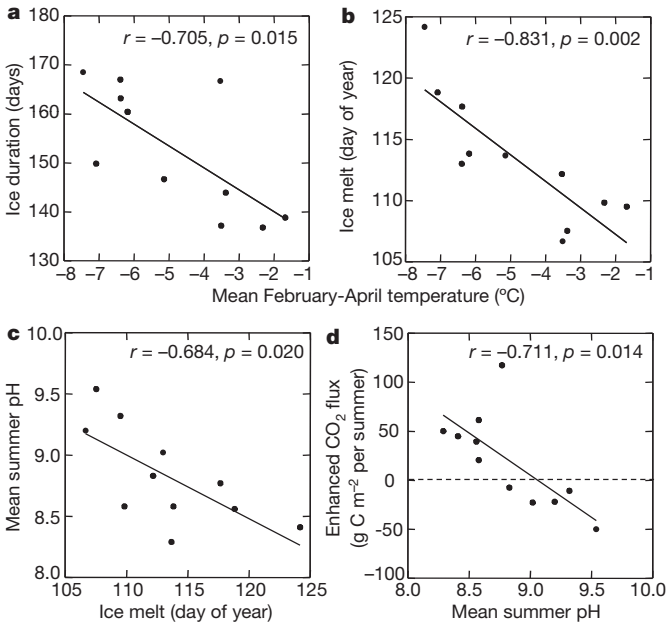


Figure 2 | Effects of winter temperature on ice cover, lake chemistry and CO_2 flux in lakes of central Canada. **a**, **b**, Least-squares regression analysis of meteorological and lake variables showing correlation of warmer winter temperatures (mean daily °C during February to April) with decreased duration of ice cover (**a**) and earlier date of ice melting the following spring (**b**). **c**, Correlation of date of ice melt with summer pH. **d**, Correlation of summer pH with CO_2 flux from hardwater lakes during summer. Regression analyses were as in Fig. 1, using 11 years with complete data.

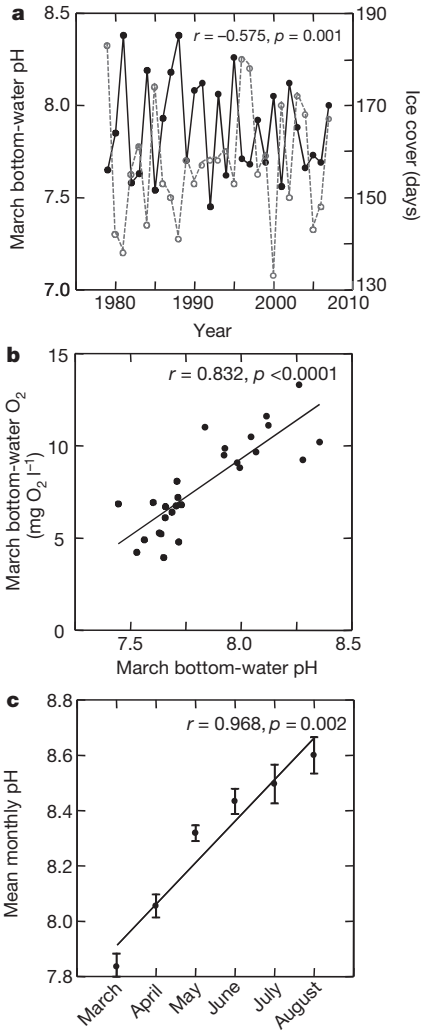


Figure 3 | Relationship between duration of ice cover, water-column pH and oxygen content before ice melt in Buffalo Pound Lake. **a**, Pearson correlation analysis showing correlation between shorter ice cover (dashed grey line) ($r = -0.575$, $P = 0.001$) and higher pH (solid black line) in water collected 1.5 m above the lake bottom during March, the month immediately before ice melt[19] ($n = 29$). **b**, Positive correlation between bottom-water pH ($r = 0.832$, $P < 0.0001$) and the oxygen content of bottom waters (mg O_2 l^{-1}) since 1979 ($n = 27$), consistent with metabolic control of lake-water pH. **c**, Linear increase in monthly pH (means ± s.e.m.; $n = 23$) with time during the ice-free season ($r = 0.968$, $P = 0.002$); the rate did not vary significantly between years ($P > 0.05$) or lakes (except Wascana). The slight nonlinearity at pH ~8.3 reflects the minimum buffering intensity of the bicarbonate–carbonate system[17]. Together, these findings suggest that longer ice cover favours metabolic production of CO_2 and depression of pH during spring and the following summer.

produces large-scale changes in CO_2 flux that are highly correlated and synchronous with those observed in the Qu'Appelle lakes[9].

Spatially coherent increases in lake water pH and CO_2 uptake imply that reduced ice cover due to atmospheric heating may create a substantial sink for atmospheric CO_2 in this large agricultural region of Canada, both by increasing pH and by extending the ice-free season for CO_2 capture. Regional winter temperature has increased by ~2.5 °C since the 1900s, resulting in a current annual mean temperature of ~2 °C and an estimated 50-day decline in ice cover as a result of earlier ice melt[23,26]. In addition, this continental climate is characterized by high inter-decadal variability[18,19], such that the maximum duration of ice cover on Qu'Appelle lakes has also declined by more than 15 days since 1980 (Fig. 3a). Analysis of land cover using the Saskatchewan Water Security Agency geographic information system with a resolution of 1:50,000 (or 1:250,000) provides documentation that permanent and largely hardwater lakes cover 11,500 km^2 (8,367 km^2 at the coarser resolution) of the 236,000 km^2 study area. Assuming that all basins have experienced a 100 g C m^{-2} per summer decline in CO_2 efflux during the past 15 years (Fig. 1d), we estimate that regional hardwater lakes may capture 1.15 Mt (0.84 Mt) more C per summer than they did during the mid-1990s, and note that this value is equivalent to 34% (25%) of present agricultural CO_2 emissions[30]. Although we recognize that such simple up-scaling has its limitations, we note that our calculations are likely to be highly conservative because they do not include estimates of CO_2 capture for September and October, months when pH remains elevated and lakes are normally free of ice.

Lakes are significant components of the global carbon budget, but show high variation in both the direction and magnitude of C fluxes and their response to global climate change[4]. Although further research is required for an evaluation of the significance of DIC-rich lakes in the global C cycle (Extended Data Fig. 6), we note that long-term monitoring of the northern Caspian Sea, a hardwater site that accounts for >40% of global inland waters, reveals a decline in ice cover similar to that observed here[31], as well as an increase in pH to nearly 9.0 (ref. 32) from earlier and lower levels[33]. Instead, our study provides the first evidence that atmospheric warming during winter has the potential to increase the pH of hardwater lakes synchronously in a large geographic region, greatly increase the rate of CO_2 sequestration by hard waters, and partly offset agricultural emissions at the subcontinental scale. Further, this work shows that global warming does not invariably increase CO_2 emissions from aquatic ecosystems[5].

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 2 September 2013; accepted 9 December 2014.
Published online 25 February 2015.

1. Burgermeister, J. Missing carbon mystery: case solved? *Nature Clim. Change* **3,** 36–37 (2007).
2. Downing, J. A. *et al.* Sediment organic carbon burial in agriculturally eutrophic impoundments over the last century. *Glob. Biogeochem. Cycles* **22,** GB1018 10.1029/2006GB002854 (2008).
3. Cole, J. J. *et al.* Plumbing the global carbon cycle: integrating inland waters into the terrestrial carbon budget. *Ecosystems* **10,** 172–185 (2007).
4. Tranvik, L. J. *et al.* Lakes and impoundments as regulators of carbon cycling and climate. *Limnol. Oceanogr.* **54,** 2298–2314 (2009).
5. Marotta, H. *et al.* Greenhouse gas production in low-latitude lake sediments responds strongly to warming. *Nature Clim. Change* **4,** 467–470 (2014).
6. Rantakari, M. & Kortelainen, P. Interannual variation and climatic regulation of the CO_2 emission from large boreal lakes. *Glob. Change Biol.* **11,** 1368–1380 (2005).
7. Sobek, S., Tranvik, L. J., Prairie, Y. T., Kortelainen, P. & Cole, J. J. Patterns and regulation of dissolved organic carbon: an analysis of 7,500 widely distributed lakes. *Limnol. Oceanogr.* **52,** 1208–1219 (2007).
8. Duarte, C. M. *et al.* CO_2 emissions from saline lakes: a global estimate of a surprisingly large flux. *J. Geophys. Res.* **113,** G04041 10.1029/2007JG000637 (2008).
9. Finlay, K., Leavitt, P. R., Patoine, A. & Wissel, B. Magnitudes and controls of organic and inorganic carbon flux through a chain of hard-water lakes on the northern Great Plains. *Limnol. Oceanogr.* **55,** 1551–1564 (2010).
10. Finlay, K., Leavitt, P. R., Wissel, B. & Prairie, Y. T. Regulation of spatial and temporal variability of carbon flux in six hard-water lakes of the northern Great Plains. *Limnol. Oceanogr.* **54,** 2553–2564 (2009).
11. del Giorgio, P. A., Cole, J. J. & Cimbleris, A. Respiration rates in bacteria exceed phytoplankton production in unproductive aquatic ecosystems. *Nature* **385,** 148–151 (1997).
12. Ask, J. *et al.* Whole-lake estimates of carbon flux through algae and bacteria in benthic and pelagic habitats of clearwater lakes. *Ecology* **90,** 1923–1932 (2009).
13. Prairie, Y. T., Bird, D. F. & Cole, J. J. The summer metabolic balance in the epilimnion of southeastern Quebec lakes. *Limnol. Oceanogr.* **47,** 316–321 (2002).
14. Bastviken, D., Tranvik, L. J., Downing, J. A., Crill, P. A. & Enrich-Prast, A. Freshwater methane emissions offset the continental carbon sink. *Science* **331,** 50 (2011).
15. Schindler, D. E., Carpenter, S. R., Cole, J. J., Kitchell, J. F. & Pace, M. Influence of food web structure on carbon exchange between lakes and the atmosphere. *Science* **277,** 248–250 (1997).
16. Hammer, U. T. *Saline Lake Ecosystems of the World* (Junk, 1986).
17. Stumm, W. & Morgan, J. J. *Aquatic Chemistry: Chemical Equilibria and Rates in Natural Waters* (Wiley, 1996).
18. Pham, S. V., Leavitt, P. R., McGowan, S., Wissel, B. & Wassenaar, L. Spatial and temporal variability of prairie lake hydrology as revealed using stable isotopes of hydrogen and oxygen. *Limnol. Oceanogr.* **54,** 101–118 (2009).

19. Vogt, R. J., Rusak, J. A., Patoine, A. & Leavitt, P. R. Differential effects of energy and mass influx on the landscape synchrony of lake ecosystems. *Ecology* **92,** 1104–1114 (2011).
20. Falkowski, P. G. *et al.* The global carbon cycle: a test of our knowledge of earth as a system. *Science* **290,** 291–296 (2000).
21. Patoine, A., Graham, M. D. & Leavitt, P. R. Spatial variation of nitrogen fixation in lakes of the northern Great Plains. *Limnol. Oceanogr.* **51,** 1665–1677 (2006).
22. Leavitt, P. R. *et al.* Paleolimnological evidence of the effects on lakes of energy and mass transfer from climate and humans. *Limnol. Oceanogr.* **54,** 2330–2348 (2009).
23. Bonsal, B. R., Prowse, T. D., Duguay, C. R. & Lacroix, M. P. Impacts of large-scale teleconnections on freshwater-ice break/freeze-up dates over Canada. *J. Hydrol. (Amst.)* **330,** 340–353 (2006).
24. Striegl, R. G. & Michmerhuizen, C. M. Hydrologic influence on methane and CO_2 dynamics in two north-central Minnesota lakes. *Limnol. Oceanogr.* **43,** 1519–1529 (1998).
25. Striegl, R. G. *et al.* Carbon dioxide partial pressure and ^{13}C content of north temperate and boreal lakes at spring ice melt. *Limnol. Oceanogr.* **46,** 941–945 (2001).
26. Magnuson, J. J. *et al.* Historical trends in lake and river ice cover in the Northern Hemisphere. *Science* **289,** 1743–1746 (2000).
27. Baehr, M. M. & DeGrandepre, M. D. Under-ice CO_2 and O_2 variability in a freshwater lake. *Biogeochemistry* **61,** 95–113 (2002).
28. Kratz, T. K., Cook, R. B., Bowser, C. J. & Brezonik, P. L. Winter and spring pH depressions in northern Wisconsin lakes caused by increase in pCO_2. *Can. J. Fish. Aquat. Sci.* **44,** 1082–1088 (2007).
29. Wissel, B., Cooper, R. N., Leavitt, P. R. & Pham, S. V. Hierarchical regulation of pelagic invertebrates in lakes of the northern Great Plains: a novel model for inter-decadal effects of future climate change. *Glob. Change Biol.* **17,** 172–185 (2011).
30. Province of Saskatchewan. *Saskatchewan's State of the Environment Report* (Government of Saskatchewan, 2008).
31. Kouraev, A. V. *et al.* Sea ice cover in Caspian and Aral Seas from historical and satellite data. *J. Mar. Syst.* **47,** 89–100 (2004).
32. Lukashin, V. N. *et al.* An integrated study in the Northern Caspian Sea: the 30th voyage of the R/V *Rift. Oceanology (Mosc.)* **50,** 439–443 (2010).
33. Tuzhilkin, V. S., Katunin, D. N. & Nalbandov, Y. R. in *Handbook of Environmental Chemistry* Vol. 5P (eds Kostianoy, A. G. & Kosarev, A. N.) 83–108 (Springer, 2005).

Acknowledgements We thank I. Phillips for estimates of lake area in Saskatchewan; D. Conrad for water chemistry data from Buffalo Pound Lake, Katherine Miller for compilation and analysis of Buffalo Pound data; J. Piwowar for assistance with GIS data; S. Pham and the University of Regina Limnology Laboratory for field work in 1996–2010; and Y. T. Prairie, E. G. Stets, J. A. Downing and D. E. Schindler for reviewing the manuscript. NSERC Canada, the Canada Foundation for Innovation, the Province of Saskatchewan, and the Canada Research Chair programme provided funding for this work.

Author Contributions P.L. and K.F. designed the study. P.L. provided data from Qu'Appelle lakes. P.L. and B.W. provided data from other hardwater lakes. B.T. conducted geochemical modelling, G.S. conducted elastic net analysis, and all authors contributed additional numerical analysis. P.L., K.F. and R.V. wrote the manuscript. All authors edited the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to P.L. (peter.leavitt@uregina.ca).

METHODS

Depth-integrated samples were collected every two weeks at standardized times and locations between May and August during 1995–2010, except at Wascana Lake (1996–2010), as part of the Qu'Appelle long-term ecological research programme[10,19,21]. For comparison, an additional 21 closed-basin, hardwater and saline lakes were sampled three to five times each summer during 2002–2010 (except 2006) using the same methods[10,18,29]. Finally, water chemistry parameters in Buffalo Pound Lake have been analysed weekly since 1979. Concentration (mmol C m^{-3}), partial pressure (p_{CO_2}), and chemically enhanced flux of CO_2 (mmol C m^{-2} d^{-1}) in Qu'Appelle lakes was calculated on each sampling date and interpolated to estimate mean summer conditions as described in refs 9 and 10. Changes in CO_2-related parameters were compared with environmental variables known to influence water chemistry, production and CO_2 flux, using unreplicated linear regression performed with SYSTAT version 10. Causes and correlates of pH decline during winter were modelled for Buffalo Pound Lake by using Geochemists Workbench version 9.0.9 and elastic net analysis in R, respectively. No statistical methods were used to predetermine sample size.

Study sites. Six study lakes are located within the Qu'Appelle River catchment, a lotic system that drains 52,000 km^2 in southern Saskatchewan, Canada (50° 00′–51° 30′ N, 101° 30′–107° 10′ W)[10,19,21] (Extended Data Fig. 1). Sites range from mesotrophic upstream lakes (Diefenbaker, Buffalo Pound and Last Mountain lakes) to eutrophic downstream sites (Katepwa and Crooked lakes), with hypereutrophic Wascana Lake located in the City of Regina (Extended Data Table 1). Dissolved inorganic (DIC) and organic carbon (DOC) concentrations are high (30–60 and 5–16 mg l^{-1}, respectively) and tend to increase with distance from headwaters, with the exception of subsaline Last Mountain Lake, which had elevated levels of both DIC and DOC. In general, lakes have moderate to high flushing rates (water residence time <1.5 years) except Last Mountain Lake (~12 years), whereas all basins show high conductivity (400–1800 μS cm^{-1}) and have elevated mean summer pH (catchment mean pH = 8.8). Lakes are polymictic in most years, except occasionally dimictic Katepwa Lake.

Estimates of pH, p_{CO_2} and chemically enhanced CO_2 flux in Qu'Appelle lakes were compared with those determined for a series of 21 hardwater and saline basins spanning an additional area of 100,000 km^2 within southern Saskatchewan[9,18,29,34]. In all cases, survey lakes lacked visible surface-water inflow and showed elevated pH (8.4–9.3) and carbon content (DOC = 10–159 mg C l^{-1}; DIC = 18–500 mg C l^{-1}), but differed greatly in size (area = 0.5–60.0 km^2; maximum depth = 1.3–30 m), mean nutrient concentrations (orthophosphate 9–610 μg l^{-1}) and salinity (total dissolved solids 0.4–50.7 g l^{-1}). Comparison of mean and variance of main chemical parameters revealed that seasonally sampled lakes were representative of a further 50 closed-basin lakes within southern Saskatchewan[18,34].

Limnological sampling. Depth-integrated samples were collected every two weeks at standardized times and locations between 1 May and 31 August during 1995–2010, except Wascana Lake (1996–2010), for a comprehensive suite of limnological variables including dissolved C species, pH, nutrient content, conductivity, O_2 content, and plankton abundance, production and composition[10,19,21]. In contrast, the 21 hardwater and saline lakes lacking surface water efflux (closed basin) were sampled three to five times each summer with standard methods during 2002–2010 (except 2006) to quantify the degree to which the Qu'Appelle lakes represented the broader prairie landscape[10,18,29]. Finally, the chemistry of water obtained from 1.5 m above the bottom of Buffalo Pound Lake (3 m depth at the sampling location) has been analysed using standard methods[35] since 1979, with comprehensive water chemistry analyses conducted at weekly intervals from 1985 to 2003.

Carbon fluxes and regulation. Concentration (mmol C m^{-3}), partial pressure (p_{CO_2}) and chemically enhanced flux of CO_2 (mmol C m^{-2} d^{-1}) were calculated on each sampling date from depth-integrated DIC concentrations (mg C l^{-1}), surface-water pH and observed wind speed (m s^{-1}) after correction for ionic strength and water temperature by using equations for both freshwater[17] and saline ecosystems[36], as detailed in refs 9 and 10. In brief, p_{CO_2} (Pa) was estimated by using Henry's Law constant and accounted for changes in temperature[37]. Chemically enhanced CO_2 flux was calculated for each sampling date in accordance with the boundary layer equations in ref. 38:

$$\text{net daily } CO_2 \text{ flux} = \alpha k\left([CO_2]_{\text{lake}} - [CO_2]_{\text{sat}}\right)$$

where $[CO_2]_{\text{lake}}$ is the concentration of CO_2 in the surface water (μmol l^{-1}), $[CO_2]_{\text{sat}}$ is the concentration of CO_2 at equilibrium with the atmosphere (μmol l^{-1}), α is the chemical enhancement of CO_2 flux at high pH[39] and was calculated from the equations in ref. 40, and k is piston velocity (cm h^{-1}) determined from equation (5) in ref. 38 relating k to wind speed, and accounting for temperature[41]. Mean wind speeds at each lake were calculated by averaging observations over all sampling dates because there were no significant differences in wind speed by month or year at a given site; they varied from 2.8 ± 2.0 m s^{-1} (Wascana Lake) to 4.3 ± 2.7 m s^{-1} (Last Mountain)[10].

Mean summer p_{CO_2} and total chemically enhanced CO_2 flux (g C m^{-2} per summer) were estimated for each lake and year by interpolation of two-weekly or monthly measurements[9,10] and integration over the summer without weighting values for differences in lake area (no substantial effect). Time series of environmental variables known to influence water chemistry, production and CO_2 flux in hardwater and saline lakes were compared with those of catchment means by using unreplicated linear regression to identify potential mechanisms regulating inter-decadal variation in C flux. Predictor time series were obtained from relevant local, regional and national sources and included air temperature, irradiance, precipitation, evaporation, river discharge and ice cover[9,10,19] as well as global climate indices including the Southern Oscillation Index, the Pacific Decadal Oscillation and the North Atlantic Oscillation. In all cases there was no significant temporal autocorrelation within annually resolved time series. Cross-correlation analysis of untransformed variables was used to determine that there were no significant lagged relationships between time series. All regression analyses were performed with SYSTAT version 10.

Elastic net analysis. Elastic net analysis[42] was used to identify and rank predictors of changes in under-ice pH in Buffalo Pound Lake during winter in the period 1985–2003. Water chemistry was analysed weekly with standard procedures[35] for samples collected from 1.5 m above the lake bottom at the 3.0-m-deep site between the date of ice-cover formation and the date of ice melt. Water parameters included dissolved oxygen (O_2), sodium (Na^+), carbonate (CO_3^{2-}), $\log_e$-transformed dissolved aluminium ($\log_e$Al), fluoride (F^-), temperature (temp), potassium (K^+), $\log_e$-transformed orthophosphate ($\log_e PO_4^{3-}$), calcium (Ca^{2+}), dissolved magnesium (Mg^+), $\log_e$-transformed nitrite + nitrate ($\log_e NO_3^-$), $\log_e$-transformed dissolved manganese ($\log_e$Mn), bromide (Br^-), total phosphorus (TP), dissolved iron (Fe), chloride (Cl^-), $\log_e$-transformed ammonium ($\log_e NH_4^+$), bicarbonate (HCO_3^-) and sulphate (SO_4^{2-}). Dates of permanent ice formation and melt were obtained from ref. 19 and from unpublished records of P.R.L.

We chose the elastic net analysis over an ordinary least-squares (OLS) multiple regression approach because OLS solutions are highly sensitive to small changes in input data (such as transformations or covariates) and because OLS imposes a hard threshold on the size of model coefficients (included or excluded). Although partial least-squares (PLS) regression can be used in place of OLS, especially when there are many correlated covariates to select from, we also preferred elastic net analysis over PLS because the former will generate sparse or parsimonious models that include only substantial predictors, whereas PLS models commonly include contributions from all covariates, irrespective of magnitude. A secondary advantage of the elastic net is that it provides only one coefficient per predictor, thereby simplifying model interpretation relative to PLS analysis.

The elastic net[42] is a regression method that minimizes a penalized deviance criterion, which places restrictions on the size of the regression coefficients, in terms of both squared and absolute values. In our application, residual sum of squares was used as the measure of deviance. The elastic net penalty takes the form

$$\lambda \sum_{j=1}^{p} (\alpha|\beta_j| + (1-\alpha)\beta_j^2)$$

where β_j is the regression coefficient for the jth covariate and α is a mixing parameter, which controls the relative weighting of the lasso (absolute) and ridge (squared) contributions and which we set to 0.5. The first term in the summation aims at producing a sparse solution with some $\beta_j = 0$; the second term handles highly correlated variables by averaging their coefficients[43]. In this formulation, λ controls the amount of shrinkage applied to the β_j. If $\lambda = 0$, the effect of the elastic net penalty is cancelled and the β_j take their full least-squares regression solutions. As λ increases from zero, the coefficients are progressively shrunk until as $\lambda \to \infty$ only the constant term (the model intercept) remains in the model. k-fold cross-validation was used to find an optimal value for λ, and we chose the simplest model (that is, the largest value of λ) within one standard error of the model with lowest cross-validation mean squared error in the interests of parsimony. The shrinkage of the coefficients implies bias in their estimated values, which we trade off with a reduction in model variance arising from the sparsity of the solution and the handling of collinear covariates. To fit the elastic net model, covariates were standardized to zero mean, unit variance. Hence the absolute size of the estimated elastic net coefficients gives an indication of the relative predictive importance for pH of each covariate. The elastic net model was fitted with the glmnet package (version 1.9-8)[44] for R (version 3.1.1 r66455)[45].

We present the results of the elastic net process via the entire path diagram of the coefficients (Extended Data Fig. 3). The path diagram shows the L1 norm (sum of the absolute values of β_j) for models along the path from the full OLS solution on the right to a model containing just a constant term on the left. Consequently, the value of λ increases from right to left in the plot. The y axis on the path diagram indicates the value of the standardized estimate for the β_j. The most parsimonious

model (the simplest model within one standard error of the best model) is indicated by the dashed line. An indication of the model complexity (degrees of freedom) is shown on the upper margin of the figure.

Geochemical modelling. Geochemical modelling was used to evaluate the importance of changes in under-ice CO_2 content on the pH of Buffalo Pound Lake for the period in which comprehensive chemical data were obtained at weekly intervals (1985–2003). As noted above, water was collected continuously from 1.5 m above the lake bottom by the Buffalo Pound Water Treatment Plant and analysed with standard protocols[35] for concentrations of all major ions, as well as dissolved oxygen (O_2), oxygen saturation (%), pH, temperature, turbidity, alkalinity, total dissolved solids, hardness, silica, orthophosphate, total phosphorus, aluminium (dissolved, particulate and total Al), dissolved and total magnesium, dissolved and total iron, nitrite + nitrate, ammonium, total nitrogen, dissolved and total manganese, bromide, fluoride, dissolved organic nitrogen, dissolved organic carbon and selected microbial parameters.

Calculations of water-column chemistry response to CO_2 production and efflux were performed with Geochemist's Workbench version 9.0.9 outfitted with the standard thermo.v8.r6+.dat database and the extended Debye–Hückel model for aqueous species activity coefficients[46–49]. Mineral precipitation was suppressed in all calculations, which is common practice for the low temperatures encountered here[17]. This decision is further justified by the fact that most of the samples in this study are supersaturated with respect to $CaCO_3$ and are therefore not governed by equilibrium with respect to this phase.

Two types of calculation were employed to explain the relationship between under-ice CO_2 production and decreases in solution pH. First, CO_2 was computationally added to the *in situ* aqueous system observed in the autumn until the pH declined to the minimum value observed in the spring. Second, actual CO_2 production under ice was estimated from declines in measured under-ice O_2 concentrations assuming a respiratory quotient of 1.2 (ref. 50). In both cases, calculations were run for each winter season using chemical parameters recorded at the date of complete ice cover in fall and the date of complete ice melt in spring. In addition, effects of CO_2 were calculated for intervals of equal ice-cover duration, but offset from the date of ice formation by one week in either direction (ice on + 1 week, ice on − 1 week). For each year, estimates of CO_2 effects were calculated as the mean of the three sets of calculations, whereas overall effects were estimated as the mean ± s.e.m. for all years in which complete water chemistry was available (usually $n = 17$). Finally, these calculations were repeated for the interval between ice formation and the pH minimum observed *in situ*, usually in mid-March. In all situations, measured solution chemistry was inputted directly into the Geochemist's Workbench React module, and the effect of CO_2 additions on pH were determined by incrementally adding CO_2 to the modelled solution and iteratively calculating the distribution of aqueous species until they met constraints imposed on the system by charge and mass balance and aqueous species equilibrium constants.

Effects of ice melt on spring CO_2 efflux and lake-water pH were also calculated using Geochemist's Workbench by incrementally subtracting CO_2 lost to the atmosphere from the aqueous solutions observed at the time of ice melt. In this case, annual time series were each standardized to the date of complete ice melt (week = 0), and changes in aqueous chemistry were analysed at weekly intervals from 1 month before ice melt (week = −4) to one month after ice melt (week = +4). Here observed declines in concentrations of total inorganic carbon (TIC) and its predominant anion (HCO_3^-) during spring were assumed to result from the interaction of several competing processes including CO_2 evasion, precipitation of $CaCO_3$, influx of inorganic C in runoff and dilution of TIC by meltwater. Unique effects of dilution were estimated from an analysis of the percentage decline in concentration of conservative ions, chloride (Cl^-) and fluoride (F^-) during the two weeks after ice melt, whereas precipitation of $CaCO_3$ was evaluated on the basis of changes in concentration of CO_3^{2-} and temporal variation in the $CaCO_3$ saturation index. After allowing for these processes, residual decline in TIC concentration was used as an estimate of the maximum possible loss of C to the atmosphere. This calculated CO_2 loss was incrementally subtracted from the fluid chemistry observed at the date of ice melt (week = 0) and its effect on lake-water pH was evaluated by comparing the observed and calculated pH changes during the two weeks after ice melt. Estimates of the magnitude of CO_2-induced pH change and unreplicated linear regression were used to evaluate the importance of vernal CO_2 efflux in potentially decoupling spring and summer pH.

Changes in partial pressure of CO_2 (p_{CO_2}) and $CaCO_3$ saturation index (Ω_{CaCO_3}) were calculated for the two-month interval bracketing the date of ice melt by inputting fluid chemistry recorded at 1.5 m depth for all relevant samples and solving the distribution of species, as discussed above. p_{CO_2} was calculated according to

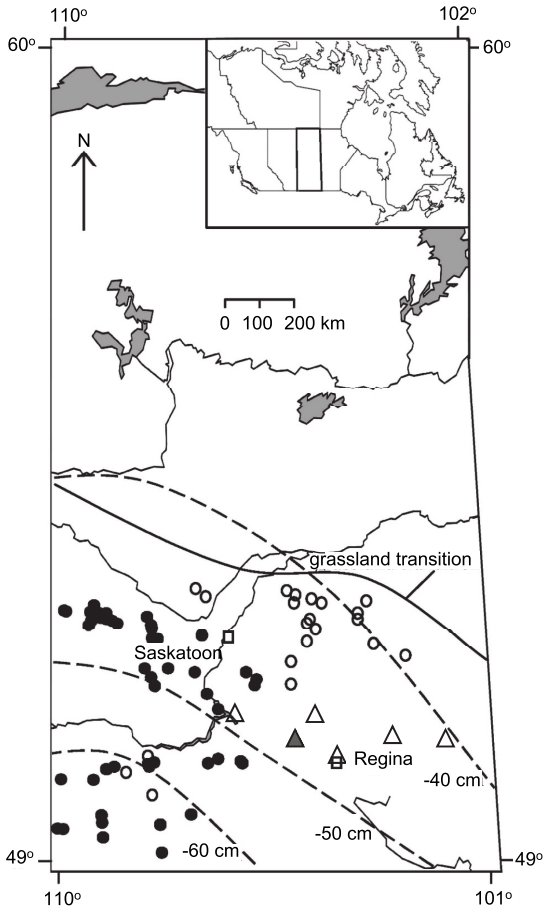
$$p_{CO_2} = \frac{a_{CO_2(aq.)}}{K_{CO_2}}$$

where $a_{CO_2(aq.)}$ is the activity of CO_2(aq.) in solution (assumed to be equivalent to its molality) and K_{CO_2} is the Henry's Law constant for CO_2 dissolution. Ω_{CaCO_3} was calculated according to

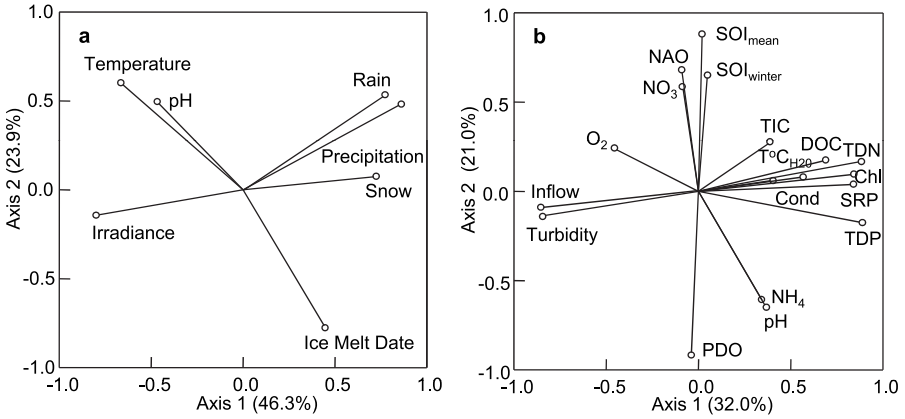
$$\Omega_{CaCO_3} = \frac{a_{Ca^{2+}} a_{CO_3^{2-}}}{K_{calcite}}$$

where $a_{Ca^{2+}}$ and $a_{CO_3^{2-}}$ are the activities of Ca^{2+} and CO_3^{2-}, respectively, calculated by multiplying the iteratively determined species molality by its activity coefficient, and $K_{calcite}$ is the equilibrium constant for calcite dissolution obtained from the thermodynamic database. Estimates of p_{CO_2} derived from Geochemist's Workbench were very highly correlated ($r^2 = 0.998$, $P < 0.0001$) with those obtained using protocols in refs 9 and 10.

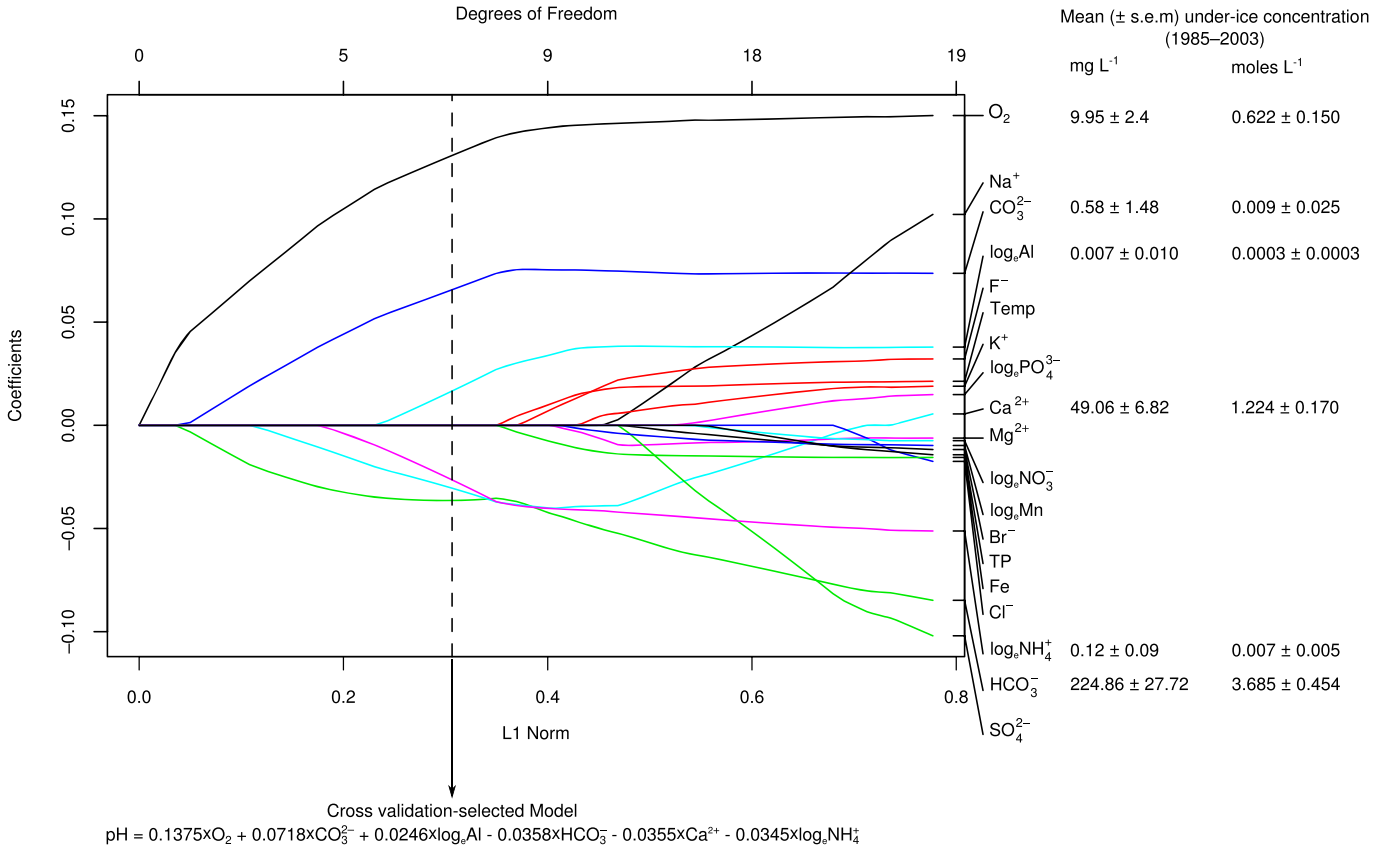
34. Pham, S. V., Leavitt, P. R., McGowan, S. & Peres-Neto, P. Spatial variability of climate and land-use effects on lakes of the northern Great Plains. *Limnol. Oceanogr.* **53,** 728–742 (2008).
35. Buffalo Pound Water Administration Board. *2012 Annual Report* (City of Regina, 2012).
36. Millero, F. J. The marine inorganic carbon cycle. *Chem. Rev.* **107,** 308–341 (2007).
37. Kling, G. W., Kipphut, G. W. & Miller, M. C. The flux of CO_2 and CH_4 from lakes and rivers in arctic Alaska. *Hydrobiologia* **240,** 23–36 (1992).
38. Cole, J. J. & Caraco, N. F. Atmospheric exchange of carbon dioxide in a low-wind oligotrophic lake measured by the addition of SF_6. *Limnol. Oceanogr.* **43,** 647–656 (1998).
39. Hoover, T. E. & Berkshire, D. C. Effects of hydration on carbon dioxide exchange across an air–water interface. *J. Geophys. Res.* **74,** 456–464 (1969).
40. Wanninkhof, R. & Knox, M. Chemical enhancement of CO_2 exchange in natural waters. *Limnol. Oceanogr.* **41,** 689–697 (1996).
41. Wanninkhof, R. Relationship between wind speed and gas exchange over the ocean. *J. Geophys. Res.* **97,** 7373–7382 (1992).
42. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. B* **67,** 301–320 (2005).
43. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning; Data Mining, Inference, and Prediction* 2nd edn (Springer, 2009).
44. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33,** 1–22 (2010).
45. R Core Team. *R: A Language and Environment for Statistical Computing* http://www.R-project.org/ (R Foundation for Statistical Computing, 2014).
46. Bethke, C. M. & Yeakel, S. *The Geochemist's Workbench Release 9.0 Reaction Modeling Guide* (Aqueous Solutions, LLC, 2013).
47. Shapley, M. D., Ito, E. & Donovan, J. J. Authigenic calcium carbonate flux in groundwater-controlled lakes: implications for lacustrine paleoclimate records. *Geochim. Cosmochim. Acta* **69,** 2517–2533 (2005).
48. Luhmann, A. J. *et al.* Experimental dissolution of dolomite by CO_2-charged brine at 100°C and 150bar: evolution of porosity, permeability, and reactive surface area. *Chem. Geol.* **380,** 145–160 (2014).
49. Tutolo, B. M., Kong, X. Z., Seyfried, W. E. Jr & Saar, M. O. Internal consistency in aqueous geochemical data revisited: applications to the aluminum system. *Geochim. Cosmochim. Acta* **133,** 216–234 (2014).
50. Berggren, M., Lapierre, J.-F. & del Giorgio, P. A. Magnitude and regulation of bacterioplankton respiratory quotient across freshwater environmental gradients. *Int. Soc. Microb. Ecol. J.* **6,** 984–993 (2012).



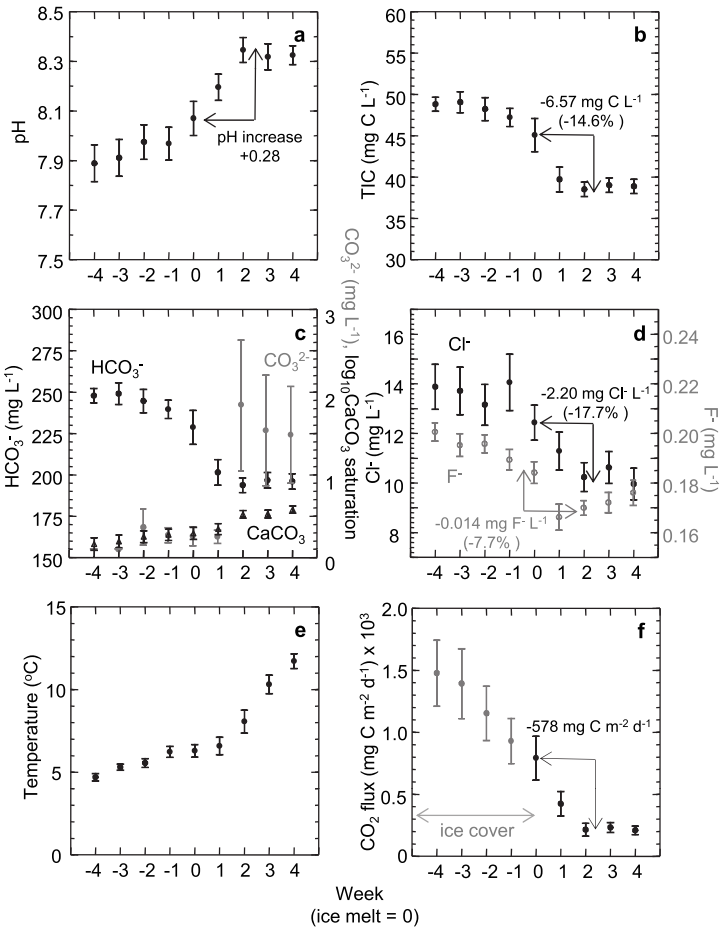
Extended Data Figure 1 | Map of study region in Saskatchewan, Canada. Hardwater lakes of the Qu'Appelle catchment (triangles) were monitored every two weeks from May to September during 1995–2010 (ref. 19), and closed-basin lakes were monitored monthly (open circles) or annually (filled circles) during 2002–2010 (except 2006)[29]. Weekly monitoring of pH occurred at Buffalo Pound Lake (black triangle) during 1979–2007. All lakes are situated in prairie grassland ecozones with pronounced precipitation deficits (annual precipitation − potential evaporation) of 40–60 cm yr^{-1} (dashed lines)[18,29].



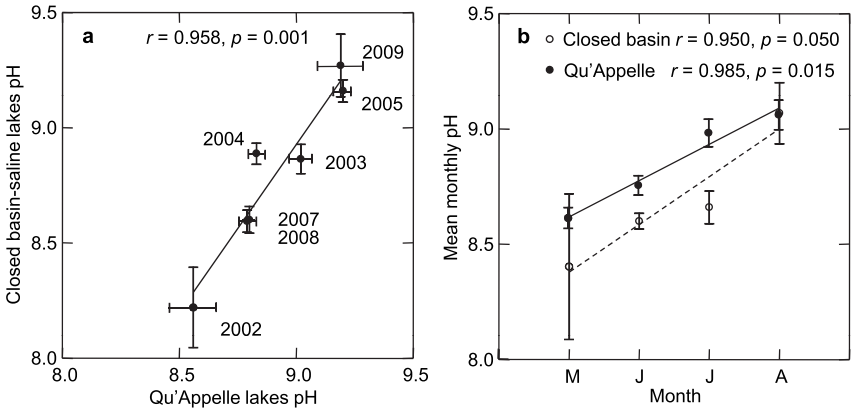
Extended Data Figure 2 | Principal components analysis of the relationship between mean annual surface water pH, annual meteorological conditions and mean summer lake parameters during 1995–2010. **a**, Ordination of mean summer pH in Qu'Appelle lakes ($n = 6$) during 1 May to 31 August in relation to mean annual meteorological conditions revealed that pH was correlated positively with mean annual and spring (not shown) temperatures, correlated negatively with the date of ice melt, and was unrelated to mean annual levels of precipitation or irradiance. Variables include $\log_{10}$-transformed mean annual temperature (temperature), total annual rainfall (rain), total annual precipitation (precipitation), total snowfall (snow), untransformed daily hours of bright sunlight (irradiance) and the calendar day of the year when ice was completely melted from the lake surface (ice melt date). **b**, Ordination of mean summer pH in relation to coeval chemical, hydrological and physical conditions in Qu'Appelle lakes, as well as indices of relevant global climate systems. Abbreviations include water temperature ($T°C_{H20}$), total inorganic carbon (TIC), dissolved organic carbon (DOC), total dissolved nitrogen (TDN), $\log_{10}$-transformed chlorophyll *a* (Chl), conductivity (Cond), $\log_{10}$-transformed soluble reactive phosphorus (SRP), $\log_{10}$-transformed total dissolved phosphorus (TDP), $\log_{10}$-transformed dissolved ammonia/ammonium (NH_4), turbidity (Secchi depth), $\log_{10}$-transformed volume of river inflow (inflow), dissolved oxygen (O_2), $\log_{10}$-transformed dissolved nitrite + nitrate (NO_3) and climate indices representing the Pacific Decadal Oscillation (PDO), the North Atlantic Oscillation (NAO) and the winter (SOI_{winter}) or annual (SOI_{mean}) Southern Oscillation Index. This analysis reveals that mean summer pH is correlated positively with the PDO and negatively with the SOI, consistent with the interpretation that warm winters and reduced ice cover result in higher summer pH in Qu'Appelle lakes.



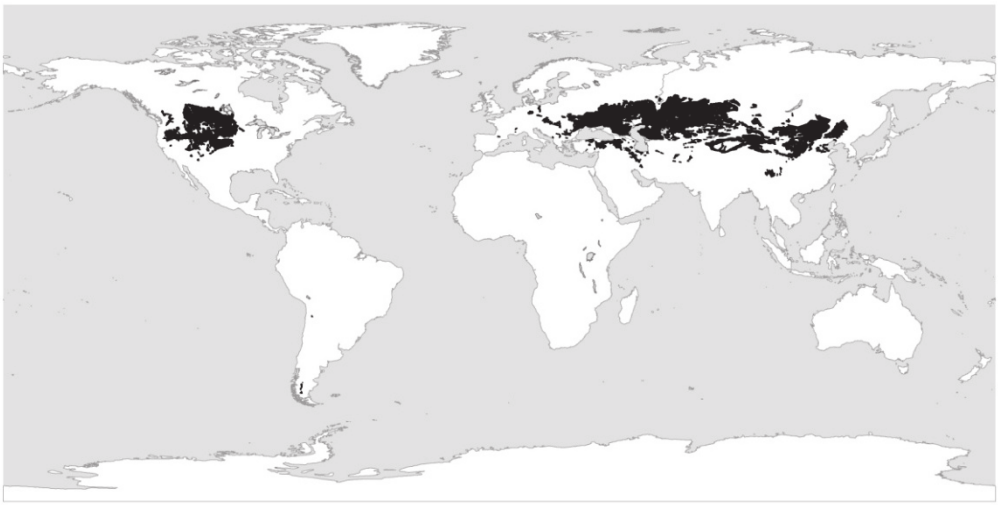
Extended Data Figure 3 | Elastic net analysis to identify and rank predictors of changes in under-ice pH in Buffalo Pound Lake during winter. Water quality parameters at 1.5 m above the lake bottom were analysed weekly using uniform methods during 1985–2003 from the date of ice-cover formation to the date of ice melt. Analysis was performed using 125 weekly observations with complete water chemistry. Parameters include concentrations of dissolved oxygen (O_2), sodium (Na^+), carbonate (CO_3^{2-}), $\log_e$-transformed dissolved aluminium ($\log_e Al$), fluoride (F^-), potassium (K^+), $\log_e$-transformed orthophosphate ($\log_e PO_4^{3-}$), calcium (Ca^{2+}), dissolved magnesium (Mg^+), $\log_e$-transformed nitite + nitrate ($\log_e NO_3^-$), $\log_e$-transformed dissolved manganese ($\log_e Mn$), bromide (Br^-), total phosphorus (TP), dissolved iron (Fe), chloride (Cl^-), $\log_e$-transformed ammonium ($\log_e NH_4^+$), bicarbonate (HCO_3^-), sulphate (SO_4^{2-}) and temperature (temp). Coloured lines indicate how standardized regression coefficients (*y* axis, left) develop (right to left) as the initial pool of predictors (*y* axis, right) is refined by removing collinear and non-significant variables. Evaluation of the standardized coefficients of the most parsimonious model (vertical dashed line; equation under graph) demonstrates that changes in microbial metabolism (O_2 decline × respiratory quotient of 1.2 = CO_2 production)[4,11] was the main factor regulating variation in water-column pH under ice, showing a nearly fourfold greater coefficient (0.14) than did either HCO_3^- or Ca^{2+} (0.04). Although dissolved $\log_e Al$, $\log_e NH_4^+$ and CO_3^{2-} were also significant predictors of changes in winter pH (standardized coefficients 0.03–0.07), concentrations of these solutes (means ± s.e.m.; $n = 17$) were too low (<0.01 M) to regulate lake-water pH relative to the effects of changes in O_2 (0.62 M), HCO_3^- (3.69 M) or Ca^{2+} (1.22 M). This analysis suggests that metabolically produced CO_2 mainly regulates variation in winter pH by the production of carbonic acid (reduces pH), but that pH decline is slightly tempered by CO_2-induced dissolution of sedimentary $CaCO_3$, the main form of sedimentary carbon in Buffalo Pound[9].



Extended Data Figure 4 | Effects of ice melt on water chemistry and spring carbon efflux from Buffalo Pound Lake, 1985–2003. **a**, Mean pH recorded at 1.5 m above the lake bottom from four weeks before to four weeks after ice melt (week = 0). The rate of pH increase was linear with time ($r = 0.96$, $P < 0.0001$) with a slightly higher magnitude of increase (0.28 units) occurring in the two weeks after ice melt. **b**, Changes in mean concentration (mg $C l^{-1}$) of total inorganic carbon (TIC) during spring. The maximum extent of TIC decline (6.57 mg $C l^{-1}$; 14.6% of TIC stock at week 0) occurred during the two weeks after ice melt as a result of a combination of CO_2 efflux (0.86 mg $C l^{-1}$) and dilution by snowmelt (5.71 mg $C l^{-1}$), as documented in **d**. **c**, Changes in concentrations of HCO_3^- and CO_3^{2-}, and $\log_{10}$ of calculated $CaCO_3$ saturation index. Patterns reveal that the decline in TIC was due to a decrease in HCO_3^- concentrations, rather than to the precipitation of $CaCO_3$. **d**, Changes in mean concentrations (mg l^{-1}) of chloride (Cl^-) and fluoride (F^-) during spring. Because concentrations of conservative tracers declined by an average of 12.7% in the two weeks after ice melt, yet TIC declined by 14.6%, we estimated that 1.9% of the decline in TIC stock at week 0 (0.86 mg $C l^{-1}$) was caused by loss of inorganic C, particularly atmospheric evasion. Geochemical modelling of water chemistry observed at week 0 demonstrated that this magnitude of CO_2 loss should increase the pH by 0.23 ± 0.08 units by week 2, a value equivalent to the observed increase in pH (see **a**). **e**, Observed changes in water temperature were too small (1.7 °C) to influence water chemistry strongly during the two weeks after ice melt. **f**, Calculated changes in chemically enhanced CO_2 efflux modelled with observed water chemistry. Ice cover was assumed to prevent potential atmospheric CO_2 exchange (grey shading), whereas most CO_2 efflux seemed to occur one to two weeks after ice melt. All water samples were collected weekly at 1.5 m above the bottom of 3.0-m-deep Buffalo Pound Lake. Error bars represent s.e.m. ($n = 17$). During each year, sampling intervals were standardized to the documented week of ice melt (week = 0) before the calculation of long-term means. Changes in $CaCO_3$ saturation and water-column p_{CO_2} were modelled with observed water-column parameters and Geochemists Workbench version 9.0.9 (see Methods).



Extended Data Figure 5 | Relationship between surface water pH in hardwater lakes of the Qu'Appelle River catchment and hydrologically closed lakes of southern Saskatchewan. **a**, Qu'Appelle lakes ($n = 6$) were monitored every two weeks during summer 1995–2010 (ref. 19); closed-basin lakes ($n = 20$ in most years) were sampled monthly or seasonally (spring, summer and autumn) during 2002–2009 (except 2006, when there were no samples)[29]. **b**, Seasonal change in mean surface water pH of 15 closed-basin lakes monitored monthly during 2002–2009. Error bars represent one s.e.m. These patterns demonstrate that the pH of closed-basin lakes varied synchronously with that of Qu'Appelle lakes on both annual and seasonal scales, despite large differences in hydrological properties[10,19].



Extended Data Figure 6 | Global map of regions where climatic conditions and soil types resemble those of southern Saskatchewan, Canada. Hardwater lake distribution is not well quantified; however, this map depicts the region in which subsoil composition favours hardwater lakes and where climatic conditions produce substantial winter ice cover. Soil data originate from the FAO–UNESCO Soil Map of the World; regions highlighted in black have subsoil concentrations of $CaCO_3$ in excess of 10% (for example Cambisol, Xerosol, Yarsol, Kastanozem and Chernozem soils). These data were overlain with temperature data (10 arcminute resolution, averaged monthly during 1950–2000) obtained from the WorldClim Global Climate Data that were restricted to regions where the monthly average temperature was below 0 °C for December–February (June–August for the Southern Hemisphere) but where the temperature was above 0 °C in October (April in the Southern Hemisphere), to exclude high-latitude lakes with permanent ice cover. The highlighted area (15,200,000 km^2) has pronounced winter and calcareous soils, spanning the prairie and steppe regions of North America, South America, Europe and Asia. If we assume this region to have a similar surface water distribution to that of southern Saskatchewan, the area occupied by permanent lakes should be between 740,678 km^2 (at 1:50,000 scale) and 538,892 km^2 (at 1:250,000). If these basins also experienced a decline in CO_2 efflux of 100 g C m^{-2} per summer during the past 15 years (Fig. 1d), global hardwater lakes may have sequestered 74.1 Mt (53.9 Mt at the coarser resolution) more C per summer than they did during the mid-1990s, a change greater than 5% of global efflux from dilute boreal lakes[3,4]. This value should increase in the future as ice cover declines.

Extended Data Table 1 | Physical and chemical characteristics of hardwater lakes of the Qu'Appelle River catchment, Saskatchewan, Canada

	Lake					
	Diefenbaker	Buffalo Pound	Last Mountain	Wascana	Katepwa	Crooked
Area (km^{-2})	500	29.1	226.6	0.5	16.2	15.0
Volume (m^3 10^6)	9400	87.5	1807.2	0.7	233.2	120.9
Mean depth (m)	33.0	3.0	7.9	1.5	14.3	8.1
WRT (yr)	1.3	0.7	12.6	0.7	1.3	0.5
TDP (μg P L^{-1})	20.9	27.8	44.9	321.6	152.3	128.8
PO_4^{3+}(μg P L^{-1})	9.7	35.4	23.6	215.8	99.7	83.3
TDN (μg N L^{-1})	421.6	511.7	990.3	1441.2	1152.6	948.4
NO_3^- (μg N L^{-1})	171.3	77.3	61.6	156.4	213.4	93.0
NH_4^+ (μg N L^{-1})	18.1	32.7	28.1	77.5	74.4	29.7
DOC (mg C L^{-1})	6.8	7.5	16.4	17.9	13.7	13.3
DIC (mg C L^{-1})	33.6	32.3	57.9	40.2	48.6	49.8
Cond (μS cm^{-1})	411.0	468.7	1776.2	900.3	1135.5	1210.7
Chl *a* (μg L^{-1})	5.4	30.9	17.2	40.8	26.2	31.3
pH	8.7	8.7	8.8	9.0	9.0	8.8

Mean summer values (May–August, 1995–2010; $n = 128$) include water residence time (WRT), total dissolved phosphorus (TDP), orthophosphate (PO_4^{3-}), total dissolved nitrogen (TDN), nitrate (NO_3^-), ammonium (NH_4^+), dissolved organic carbon (DOC), dissolved inorganic carbon (DIC), conductivity (Cond) and algal abundance as chlorophyll *a* (Chl *a*). Data are from previous papers[9,19] and unpublished records of P.R.L.

LETTER

doi:10.1038/nature13996

Spatiotemporal transcriptomics reveals the evolutionary history of the endoderm germ layer

Tamar Hashimshony[1], Martin Feder[1], Michal Levin[1], Brian K. Hall[2] & Itai Yanai[1]

The concept of germ layers has been one of the foremost organizing principles in developmental biology, classification, systematics and evolution for 150 years (refs 1–3). Of the three germ layers, the mesoderm is found in bilaterian animals but is absent in species in the phyla Cnidaria and Ctenophora, which has been taken as evidence that the mesoderm was the final germ layer to evolve[1,4,5]. The origin of the ectoderm and endoderm germ layers, however, remains unclear, with models supporting the antecedence of each as well as a simultaneous origin[4,6–9]. Here we determine the temporal and spatial components of gene expression spanning embryonic development for all *Caenorhabditis elegans* genes and use it to determine the evolutionary ages of the germ layers. The gene expression program of the mesoderm is induced after those of the ectoderm and endoderm, thus making it the last germ layer both to evolve and to develop. Strikingly, the *C. elegans* endoderm and ectoderm expression programs do not co-induce; rather the endoderm activates earlier, and this is also observed in the expression of endoderm orthologues during the embryology of the frog *Xenopus tropicalis*, the sea anemone *Nematostella vectensis* and the sponge *Amphimedon queenslandica*. Querying the phylogenetic ages of specifically expressed genes reveals that the endoderm comprises older genes. Taken together, we propose that the endoderm program dates back to the origin of multicellularity, whereas the ectoderm originated as a secondary germ layer freed from ancestral feeding functions.

Embryonic development in *C. elegans* begins with a series of asymmetric cell divisions producing five somatic founder cells (AB, MS, E, C, D), each giving rise to a limited number of tissue types, and a single germline founder cell (P4) (Fig. 1a)[10]. To determine globally the spatiotemporal gene expression in the *C. elegans* embryo, we isolated five blastomeres (AB, MS, E, C and P3) that collectively amount to the entire embryo and cultured them *in vitro*[11] to obtain a time course (Fig. 1a and Extended Data Fig. 1). The blastomeres divided well *in vitro*, maintaining the expected relative division rates: all AB cells maintained a synchronized division rate, while E divided slower than MS (Extended Data Fig. 1). We analysed the transcriptomes of these collected blastomeres using our recently described cell expression by linear amplification and sequencing (CEL-seq) method[12] for performing single-cell RNA sequencing[13,14]. To assay the degree to which the cultured blastomeres exhibit the expected expression, we also generated a whole-embryo CEL-seq time course, spanning the one-cell stage to the free-living larva, at 10 min resolution up to muscle movement, and then roughly every 30 min (Fig. 1a).

The quality of the data set was assessed in several ways. First, an average Pearson's correlation coefficient of the biological replicates of 0.9 indicates both that the blastomeres follow similar paths as they differentiate in isolation and that the CEL-seq method is reproducible (Extended Data Fig. 2a). Second, we compared the whole-embryo transcriptomes with a weighted sum of the time courses of the five lineages (Fig. 1b), and found that the blastomere data mirror the gene expression of the whole embryo, at the expected times (circles in Fig. 1b). Third, we show that the overall differentiation *in vitro* is intact, as the blastomere lineages express the expected differentiation events (Fig. 1c). Finally, we found that these profiles compared well with a previously published set of embryonic expression profiles[15] (Extended Data Fig. 2e and Supplementary Table 1). Our data reveal the spatial and temporal expression profile for each gene (Fig. 1d). For example, *unc-120*/SRF has expression in MS, C and P3, as expected from its known role as a myogenic master regulator[16].

Since the five lineages each develop in isolation from one another, their context in the embryo is lost and, consequently, absence of signalling between cell lineages must affect some gene regulation. Most noticeably, the specification of the pharynx in the AB lineage is dependent upon two Notch signalling events[17] and indeed we do not see expression of pharyngeal specification genes in the AB lineage (Extended Data Fig. 3a). Thus, although we found that for some genes expected levels are maintained (for example *wrm-1*, a β-catenin-like protein, *pal-1*/caudal and *pie-1*, a zinc-finger protein; Fig. 1d), for some genes expression is higher than in the whole embryo (*flp-15*; Fig. 1d) and for others expression is at lower levels (*ceh-27*, a homeodomain protein and Y41D4B.26; Fig. 1d). We found a general coherence between the time courses: 82% of the genes are within one $\log_2$ unit difference (Extended Data Fig. 3b). Of the genes that do differ, we found a strong bias for genes with lower expression in the blastomere time course as opposed to higher expression. For 380 genes expressed in the whole-embryo time course, we detected no expression at all in the blastomere time courses (Supplementary Table 2; see, for example, C55B7.3 in Fig. 1d). Genes with 'missing' expression tend to be expressed late in development (Extended Data Fig. 3f), indicating that, although in earlier development very few genes are unaccounted for in the data set, by the end of the time course noticeable deviations from standard development are apparent.

Performing principal component analysis on the blastomere transcriptomes distinguished the three germ layers (Fig. 2a). The three principal components collectively explained 41% of the variation in gene expression across the five lineages. The first principal component (PC1) correlated with developmental time, reflecting the expression of genes with non-specific expression (Extended Data Fig. 4). In general, PC2 distinguished the endoderm while PC3 distinguished ectoderm from mesoderm (Fig. 2a). The C lineage clusters with the other mesodermal lineages, although it produces both muscle and epidermis, probably because it contains twice as many muscle cells as epidermal cells[10]. The overall distribution of the time courses into germ layers provides evidence for their distinction at the transcriptomic level.

To identify the specific genes uniquely expressed in each germ layer, we computed the correlation of the expression profile of each of the dynamically expressed genes to all others, and clustered them using hierarchical clustering (Fig. 2b). We detected 25 clusters, each comprising at least 10 genes. Gene members in a given cluster tended to have the same timing and location of expression (Fig. 2b, see right-hand bars). Fifty-four per cent of dynamically expressed genes are not specific to particular lineages (Fig. 2b), with nearly half deriving from the maternal transcriptome. The dynamically expressed genes with lineage specificity were divided according to their germ layer of expression (Extended Data Fig. 5), while further requiring each germ-layer annotated gene to have at least two-thirds of its expression in that germ layer (Supplementary

[1]Department of Biology, Technion – Israel Institute of Technology, Haifa 32000, Israel. [2]Department of Biology, Dalhousie University, Halifax, Nova Scotia B3H 4JI, Canada.

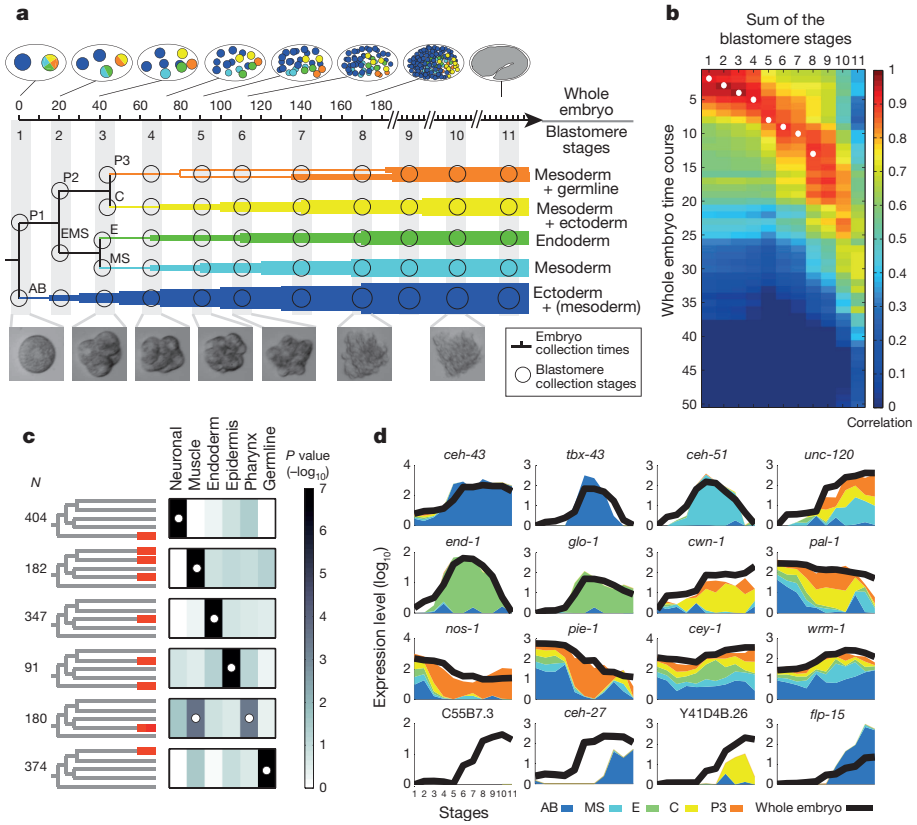


Figure 1 | Determining the expression profiles of the *C. elegans* embryonic founder cell lineages. **a**, Sample collections are indicated for the shown *C. elegans* blastomere lineages (circles, Extended Data Fig. 1 and Extended Data Tables 1 and 2) and whole embryos (notches, in minutes). **b**, Heat map showing Pearson's correlation coefficients among the transcriptomes of the whole embryo and the sum of the individual blastomere lineages. White circles indicate pairs of blastomere stages and embryonic time-points expected to be most similar (Extended Data Table 2). **c**, *P* values of enrichment across curated lists of genes for the indicated lineage-specific gene expression clusters (Extended Data Table 3 and Extended Data Fig. 2d). The white circles indicate the expected differentiation of each expression cluster (Extended Data Table 1). **d**, Spatial and temporal gene expression profiles.

Table 4) . Mapping these to their time of induction in the whole embryo, we found that germ-layer-specific expression increases with developmental time (Fig. 2c). Moreover, different germ layers initiate their programs at different times: first the endoderm, then the ectodermal expression and finally the mesodermal expression (Fig. 2c). This general pattern is also reflected when examining the dynamics of the germ layers through their average expression of the genes (Fig. 3).

The dynamics of the germ-layer expression programs may be unique to *C. elegans* or a general property of animal development. To test this, we analysed the previously characterized transcriptomes of the distantly related species *X. tropicalis*[18], *N. vectensis*[19] and *A. queenslandica*[20]. For each species, we mapped the orthologues of the *C. elegans* germ-layer genes in the respective genome and computed their average developmental expression profiles. We found a general recapitulation of

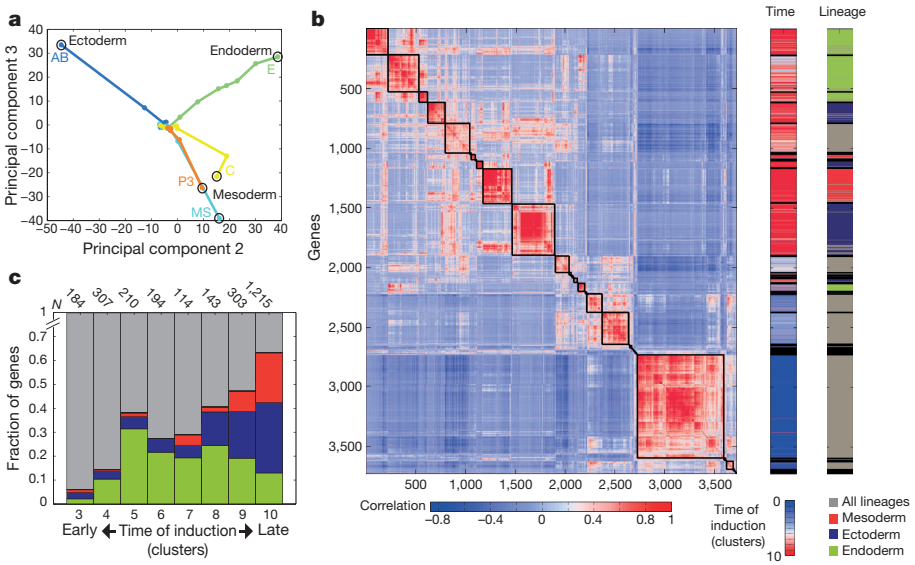


Figure 2 | Dynamics of germ-layer gene expression throughout development. **a**, Principal component analysis on dynamically expressed genes for the five lineage time courses (see Extended Data Fig. 4 for principal component 1). Adjacent stages of the same lineage are connected by a line; the terminal stage is indicated by a circle. **b**, Heat map indicating Pearson's correlation coefficients between blastomere expression profile of dynamically expressed genes. The right-hand side bars indicate the time of expression (temporal clusters, as in Extended Data Fig. 3c–e) and the location of expression. **c**, Summary of location of expression for genes according to temporal clusters.

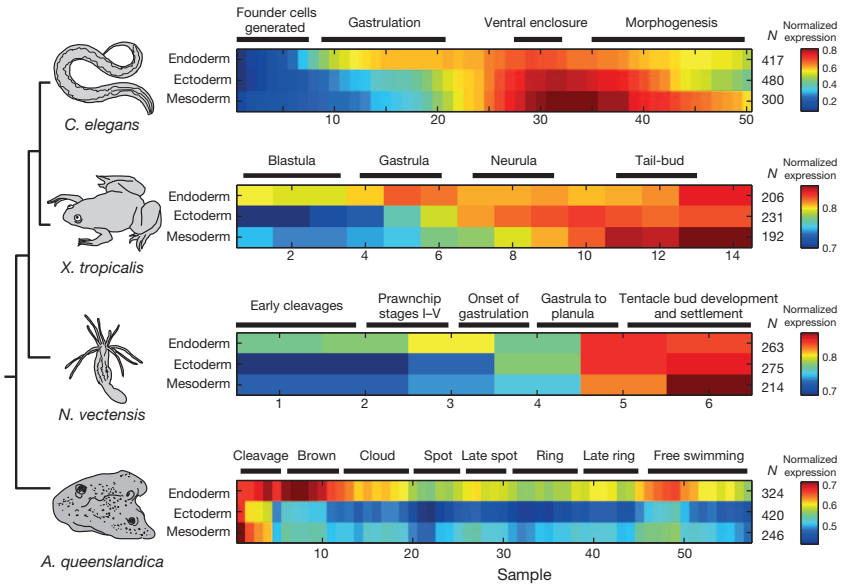


Figure 3 | The endoderm expression program precedes the ectoderm program in diverse species. Expression of germ-layer genes in *C. elegans*, and their orthologues in *X. tropicalis*, *N. vectensis* and *A. queenslandica*. The average is computed on the maximum-normalized gene profiles.

the order found in *C. elegans* (Fig. 3). The onset of the endodermal program in *Xenopus* occurs during gastrulation, well before that of the ectodermal and mesodermal programs ($P < 0.01$, Kolmogorov–Smirnov test). In *Nematostella*, we also detected a major rise in the expression of endoderm orthologues during gastrulation ($P < 10^{-3}$). The observation that mesoderm orthologues in *Nematostella* are expressed in the planula is consistent with the notion that the bilaterian mesoderm was co-opted from late-expressed genes. In *Amphimedon*, endoderm orthologues are enriched for expression during the 'brown' stage, in which two layers first become visible. Expression of the orthologues of the ectoderm and mesoderm germ-layer genes, in contrast, is seen only in the early stages ($P < 10^{-4}$), reflecting that they are solely deposited as maternal transcripts.

The distinct and conserved temporal inductions of germ-layer-specific expression (Fig. 3), with the mesoderm both appearing last in evolutionary timescales and developing last in the embryo, support accretion of processes as a mechanism in the evolution of development[7]. Extending this reasoning to the endoderm suggests that it originated before the ectoderm. According to this scenario, the endoderm is expected to express genes of older origin. To test this, we studied gene ages using the phylostratigraphy approach, which infers a gene's age from the phylogenetic breadth of its orthologues[21]. For a set of temporal stages, we computed for genes dynamically expressed at those times the fraction having orthologues in non-metazoan opisthokont eukaryotes. Using this analysis, we found that genes expressed in mid-development are generally of older origin than those expressed at other embryonic stages (Fig. 4a and Extended Data Fig. 6), consistent with previous analyses[21–23]. Examining the evolutionary age of the individual germ layers, we found that genes specifically expressed in the endoderm have a significantly higher fraction of older genes ($P < 10^{-5}$, χ^2 test). In contrast, the ectoderm and mesoderm genes are significantly younger ($P < 10^{-3}$, χ^2 test).

Since the phylogenetic analysis revealed that endoderm genes comprise genes of older origin, we enquired into their functional properties. We found that endoderm-specific genes are enriched for energy production, metabolism and transport functions (Fig. 4b and Extended Data Fig. 7). The observation that the endoderm is enriched in general feeding functions suggests that it is closer, relative to the ectoderm, in its characteristics to the choanoflagellate-like ancestor. To test this, we

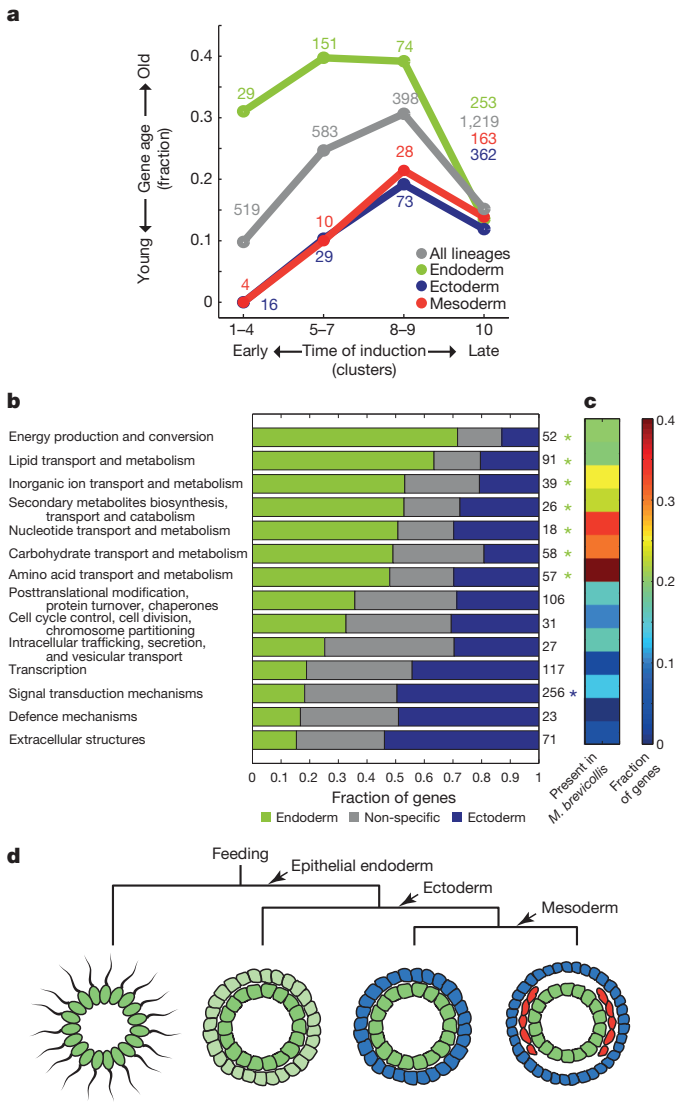


Figure 4 | The germ layers exhibit distinct gene ages and functional category enrichments. **a**, Fraction of 'old' genes—defined as presence of orthologues in other opisthokont eukaryotes—across the indicated temporal induction clusters and germ layers. Different gene age thresholds show similar results (Extended Data Fig. 6). **b**, For the functional categories shown, the bars indicate the fraction of genes in the endoderm gene set, ectoderm gene set, and other dynamic and zygotically expressed genes. Asterisks indicate significant endoderm (green) and ectoderm (blue) enrichments ($P < 0.01$, hypergeometric distribution). **c**, The fraction of orthologues in *M. brevicollis* is indicated for each functional category. **d**, A model for germ-layer evolution.

examined the level of orthology with the choanoflagellate *Monosiga brevicollis*[24] for each of the functional classes. Indeed we found a higher fraction of *M. brevicollis* orthologues in endoderm-enriched functional classes, such as transport and metabolism (Fig. 4c), suggesting that the endoderm is most closely aligned with the feeding capabilities of the free-living choanoflagellates. Moreover, while transport and metabolism appear to be related to 'housekeeping' functions, we observe, in contrast, that they are induced early on in embryogenesis in the endoderm germ-layer program.

Our results shed light on the evolutionary history of the endoderm germ layer (Fig. 4d). At the dawn of the metazoans, choanoflagellate-like colonial organisms comprised individual cells that probably all retained feeding functions. However, with the evolution of epithelial cells, the possibility of distinct cell-types emerged, as cells could communicate by strong membrane connections. Our analysis of the composition and dynamics of the germ-layer transcriptomes leads us to propose that the endoderm program has retained the feeding functions of its choanoflagellate-like ancestor. Expression in the *Amphimedon* sponge is informative since physical layers of epithelia[25] exist in this organism. The expression of sponge orthologues of the endoderm gene set suggests that *Amphimedon* only has a functional 'proto-endoderm' germ layer. This is also supported by recent evidence that the *GATA* gene in *Amphimedon* is expressed in the internal layer in the sponge[26].

In the lineage leading to the eumetazoans, the transport and metabolic functions performed by internal cells may have allowed the external cells to specialize into an ectodermal germ layer (Fig. 4d). In this model, the ancestry of the endoderm follows from its role in feeding, whereas only later in evolution was it coupled with its current function as the gastrulating internal layer. This scenario is in line with Haeckel's gastrea hypothesis[27,28] which posits a layered spherical organism as the urmetazoan. However, our model of feeding processes driving selection of the endodermal identity is also consistent with an ancestral flattened placula, as proposed by Bütschli[29,30], that subsequently evolved into a two-layered stage where the lower epithelia specialized in digestion.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 2 July; accepted 23 October 2014.
Published online 10 December 2014.

1. Hall, B. K. *Evolutionary Developmental Biology* 2nd edn (Chapman & Hall, 1998).
2. Wolpert, L. *Principles of Development* 4th edn (Oxford Univ. Press, 2011).
3. Technau, U. & Scholz, C. B. Origin and evolution of endoderm and mesoderm. *Int. J. Dev. Biol.* **47,** 531–539 (2003).
4. Ryan, J. F. *et al.* The genome of the ctenophore *Mnemiopsis leidyi* and its implications for cell type evolution. *Science* **342,** 1242592 (2013).
5. Martindale, M. Q., Pang, K. & Finnerty, J. R. Investigating the origins of triploblasty: 'mesodermal' gene expression in a diploblastic animal, the sea anemone *Nematostella vectensis* (phylum, Cnidaria; class, Anthozoa). *Development* **131,** 2463–2474 (2004).
6. Buss, L. W. *The Evolution of Individuality* (Princeton Univ. Press, 1987).
7. Gould, S. J. *Ontogeny and Phylogeny* (Belknap Press of Harvard Univ. Press, 1977).
8. Nielsen, C. *Animal Evolution: Interrelationships of the Living Phyla* 3rd edn (Oxford Univ. Press, 2012).
9. Valentine, J. W. *On the Origin of Phyla* (Univ. Chicago Press, 2004).
10. Sulston, J. E., Schierenberg, E., White, J. G. & Thomson, J. N. The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Dev. Biol.* **100,** 64–119 (1983).
11. Edgar, L. G. & Goldstein, B. Culture and manipulation of embryonic cells. *Methods Cell Biol.* **107,** 151–175 (2012).
12. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-seq: single-cell RNA-seq by multiplexed linear amplification. *Cell Rep.* **2,** 666–673 (2012).
13. Shalek, A. K. *et al.* Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498,** 236–240 (2013).
14. Shapiro, E., Biezuner, T. & Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Rev. Genet.* **14,** 618–630 (2013).
15. Murray, J. I. *et al.* Multidimensional regulation of gene expression in the *C. elegans* embryo. *Genome Res.* **22,** 1282–1294 (2012).
16. Fukushige, T., Brodigan, T. M., Schriefer, L. A., Waterston, R. H. & Krause, M. Defining the transcriptional redundancy of early bodywall muscle development in *C. elegans*: evidence for a unified theory of animal muscle development. *Genes Dev.* **20,** 3395–3406 (2006).
17. Neves, A. & Priess, J. R. The REF-1 family of bHLH transcription factors pattern *C. elegans* embryos through Notch-dependent and Notch-independent pathways. *Dev. Cell* **8,** 867–879 (2005).
18. Yanai, I., Peshkin, L., Jorgensen, P. & Kirschner, M. W. Mapping gene expression in two *Xenopus* species: evolutionary constraints and developmental flexibility. *Dev. Cell* **20,** 483–496 (2011).
19. Helm, R. R., Siebert, S., Tulin, S., Smith, J. & Dunn, C. W. Characterization of differential transcript abundance through time during *Nematostella vectensis* development. *BMC Genomics* **14,** 266 (2013).
20. Anavy, L. *et al.* BLIND ordering of large-scale transcriptomic developmental timecourses. *Development* **141,** 1161–1166 (2014).
21. Domazet-Loso, T. & Tautz, D. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* **468,** 815–818 (2010).
22. Levin, M., Hashimshony, T., Wagner, F. & Yanai, I. Developmental milestones punctuate gene expression in the *Caenorhabditis* embryo. *Dev. Cell* **22,** 1101–1108 (2012).
23. Kalinka, A. T. *et al.* Gene expression divergence recapitulates the developmental hourglass model. *Nature* **468,** 811–814 (2010).
24. King, N. *et al.* The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. *Nature* **451,** 783–788 (2008).
25. Leys, S. P. & Riesgo, A. Epithelia, an evolutionary novelty of metazoans. *J. Exp. Zool. B* **318,** 438–447 (2012).
26. Nakanishi, N., Sogabe, S. & Degnan, B. M. Evolutionary origin of gastrulation: insights from sponge development. *BMC Biol.* **12,** 26 (2014).
27. Haeckel, E. Die Gastraea-Theorie, die phylogenetische Classification des Thierreichs und die Homologie der Keimblätter. *Jenaische Z. Naturwiss.* **8,** 1–55 (1874).
28. Leininger, S. *et al.* Developmental gene expression provides clues to relationships between sponge and eumetazoan body plans. *Nature Commun.* **5,** 3905 (2014).
29. Schierwater, B. *et al.* Concatenated analysis sheds light on early metazoan evolution and fuels a modern "urmetazoon" hypothesis. *PLoS Biol.* **7,** e20 (2009).
30. Bürtschli, O. Bemerkungen zur Gastraea-Theorie. *Morph. Jahrb.* **18,** 415–427 (1884).

Supplementary Information is available in the online version of the paper.

Acknowledgements We acknowledge the contribution of computational analyses by D. H. Silver, L. Anavy and F. Wagner in an early stage of this project. We also acknowledge advice from B. Degnan, A. Cole, M. Adamska and A. Polsky. We thank the Technion Genome Center for technical assistance. This work was supported by a European Research Council grant (EvoDevoPaths) and the EMBO Young Investigator Program.

Author Contributions T.H. and I.Y. designed the experiment. T.H. performed the experiments. M.L. contributed whole-embryo data. M.F. performed the initial analysis on the RNA-sequencing data. I.Y. analysed the data with help from T.H. and M.F. T.H., B.K.H. and I.Y. wrote the manuscript.

Author Information The complete data set has been deposited in the National Center for Biotechnical Information Gene Expression Omnibus database under accession number GSE50548. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to I.Y. (yanai@technion.ac.il).

METHODS

Blastomere isolation and culturing. Egg shells were removed from *C. elegans* embryos and the resulting blastomeres cultured as previously described[11]. The egg shell and vitelline membrane were removed at the two-cell stage, and the embryo separated to the AB and P1 blastomeres by pipetting. P1 was allowed to undergo one cell division and separated to EMS and P2, or two cell divisions before being separated to the MS, E, C and P3 blastomeres, to allow the *Wnt* signalling from P2 to EMS (Extended Data Fig. 1)[31]. The five lineages were cultured in a humid chamber in EGM[11], and division of the E blastomere was used as a clock (Extended Data Table 2). All lineages from a single embryo were frozen at the same time. Individual samples were transferred with a micro-pipette into a 0.5 µl drop of egg salts placed on the cap of a 0.5 ml Lobind Eppendorf tube, excess liquid was aspirated off, and the samples frozen in liquid nitrogen. Samples were stored at −80 °C. Samples were collected in triplicates; correlations between replicates are shown in Extended Data Fig. 2a. Throughout this work, 'correlation' denotes Pearson's correlation coefficient.

Whole-embryo time course. Precisely staged single embryos were collected at the one-, two- and four-cell stages, and 10 min intervals thereafter up to muscle movement, then roughly every 30 min; 50 embryos were used in total. RNA from each embryo was prepared using TRIzol as previously described[22] with one modification: 1 µl of the ERCC spike-in kit[32] (1:500,000 dilution) was added with the TRIzol to each sample.

Single cell and whole-embryo transcriptomics. CEL-seq[12] was used to amplify and sequence both RNA from the whole embryos and the cultured blastomeres. For the whole embryos, RNA was re-suspended in 5 µl water and 1 µl primer added; 1.2 µl were taken for the amplification. For the blastomeres, 1 µl of a 1:500,000 dilution of the ERCC spike-in kit and 0.2 µl of the primer were mixed (a total of 1.2 µl) and added directly to the lid of the Eppendorf tube where the cell was frozen. Linear amplification and library preparation were as previously described[12]. Libraries were sequenced on an Illumina HiSeq2000 according to standard protocols. Paired-end sequencing was performed, reading at least 11 bases for read 1, 35 bases for read 2, and the Illumina barcode when needed. The complete data set has been deposited in the Gene Expression Omnibus database under accession number GSE50548.

Expression analysis pipeline. Transcript abundances were obtained from the sequencing data as previously described[12]. Briefly, libraries were sequenced on an Illumina HiSeq2000 according to standard, paired-end sequencing, using the CEL-seq protocol[12]. Mapping of the reads used BWA[33], version 0.6.1, against the *C. elegans* WBCel215 genome (bwa aln -n 0.04 -o 1 -e -1 -d 16 -i 5 -k 2 -M 3 -O 11 -E 4). Read counting used htseq-count version 0.5.3p1 defaults, against WS230 annotation exons. The counts were normalized by dividing by the total number of mapped reads for each gene and multiplying by 10^6, yielding the estimated gene expression levels in transcripts per million (t.p.m.).

Warped whole-embryo time course. The whole-embryo time course (Extended Data Fig. 2c) was compared with the blastomere time courses (Fig. 1b) using a restricted set of 4,527 genes with a $\log_2$ fold-change of at least 5 across the 50-embryo time course, greater than 100 t.p.m. maximum expression, and less than 10 t.p.m. minimum expression. These cutoffs were used to limit analysis to only the most dynamically expressed genes given the distinct dynamics of the whole-embryo time course. The minimum expression threshold further selected for temporally restricted expression. For each blastomere time point, the five lineages were summed up to represent the whole embryo, taking into account the fraction of the whole embryo represented by the specific lineage (half for AB, one eighth each for E, MS, C and P3). An eleven-stage warped whole-embryo time course was generated by taking for each stage a weighted average across the 50 embryos based upon the correlations with the blastomere time course, raised to the tenth power. Different definitions of this set resulted in very similar warped profiles.

Spatial and temporal gene expression profiles. In the profiles shown in Fig. 1d, the log expression is split among the lineages according to the fraction in the natural scale expression. The black line indicates the expression of the whole-embryo time course.

Definition of gene sets for dynamically expressed and differentiation genes. The 3,910 dynamically expressed genes were defined based upon the warped whole-embryo time course with >3 $\log_2$ fold-change, >10 t.p.m. maximum expression and <100 t.p.m. minimum expression (Extended Data Fig. 2b). These parameters were adapted to the warped time course, which is less dynamic owing to averaging effects. 'Constitutively expressed' genes (Extended Data Fig. 3b) were defined as highly expressed genes (>500 t.p.m. maximum expression) but not members of the dynamically expressed genes. 'Expressed genes' (Extended Data Fig. 3b) were defined as those with >10 t.p.m. maximum expression. The differentiation gene sets (Fig. 1c and Extended Data Fig. 2d) were generated for each group—neurons (AB), muscle (MS, C and P3), endoderm (E), epidermis (AB and C), pharynx (MS) and germline (P3)—by examining terminal expression in the time courses. Genes were assigned to one of the seven sets if they exhibited expression ≥ 50 t.p.m. in that group and a correlation coefficient greater than 0.7 of expression across the lineages with the expected expression pattern, as highlighted in red on the lineage trees. The parameters were set according to their definition of similarly sized sets.

Clusters of temporal gene expression patterns. A correlation coefficient was computed for each gene's temporal warped whole-embryo time course against each of 17 idealized expression profiles (Extended Data Fig. 3c). The idealized profiles were constructed based upon average expression of clusters using the *k*-means algorithm and represent the general patterns of the transcriptome. The idealized profiles are vectors of the same length (11) as the warped time-course profile but with digital expression of three possible values: 0, 1 and 2. Each dynamically expressed gene was then assigned to the idealized profile to which it best correlateds. Seven of the 17 idealized profiles correspond to 'maternal' profiles (Extended Data Fig. 3c) in which expression is initially high and then drops. We collapsed these seven profiles to one profile and denoted it as the '0' cluster in Fig. 2b.

Hierarchical clustering and definition of germ-layer genes. Hierarchical clustering used the 'linkage' function in MATLAB using the unweighted centre of mass distance (UPGMC) algorithm. The top 20 clusters with at least ten genes were examined (Fig. 2b). Clusters with at least 65% of the genes of the same germ layer contributed their genes with the dominant germ layer. Germ layers were assigned by correlating the average expression with germ-layer-specific patterns with a cutoff of 0.6 correlation with the following idealized vectors: endoderm = [00100]; ectoderm = [10000]; mesoderm = [01011], where the order is AB, MS, E, C and P3. Germ-layer genes were defined according to the sum of the genes identified by the clusters and are indicated in Fig. 2b. We further filtered the germ-layer gene sets by keeping only those genes whose expression was partitioned across the germ layers such that at least two-thirds of the expression was in that germ layer.

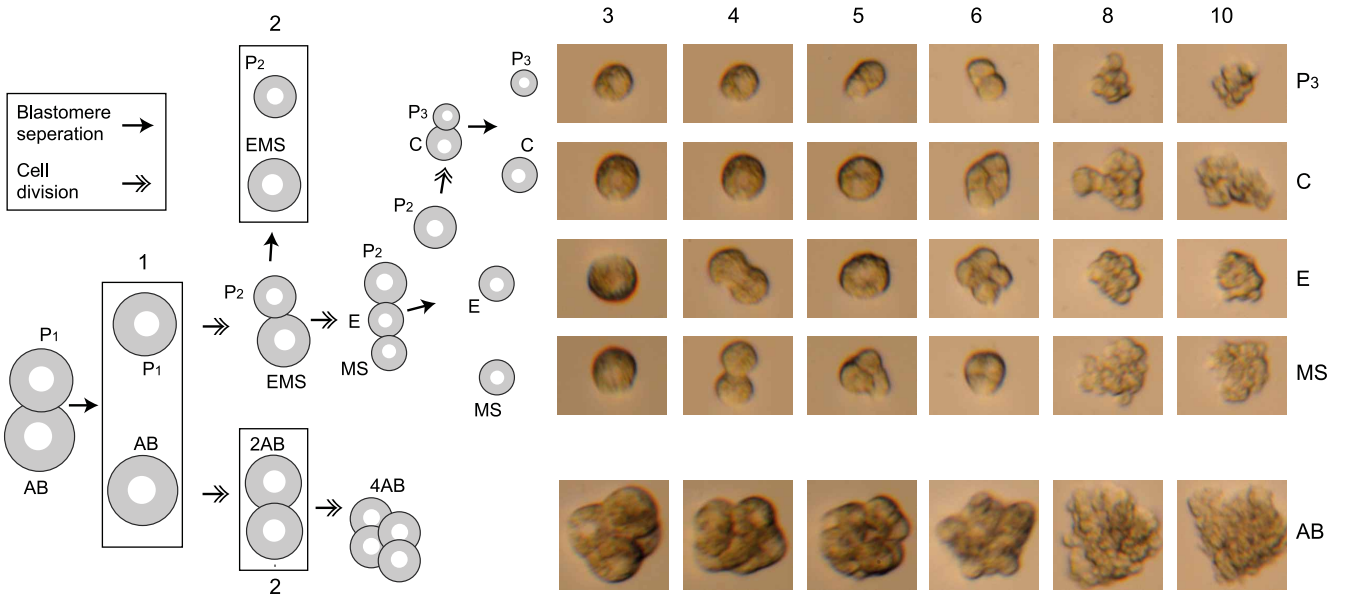
Gene age. Orthologies were retrieved from the MetaPHoRs project using the 2010 release[34]. Taxonomies were retrieved from the NCBI Taxonomy. For each *C. elegans* gene, if the gene was also present in at least 25% of the examined non-metazoan ophistokont eukaryotes, it was annotated as 'old'. Similar results were also observed for the definition of 'old' genes at the level of eukaryotes and cellular life (Extended Data Fig. 6). MetaPHoRs were also used to delineate the orthologies shown in Fig. 4c for *M. brevicollis*.

Orthologous gene expression profiles. The developmental time courses of *A. queenslandica*, *X. tropicalis* and *N. vectensis* have been previously described[18–20]. For these species, the latest protein annotations were used to detect orthologies as follows: *A. queenslandica*, Aqu2; *X. tropicalis*, JGI_4.2; *N. vectensis*, GCA_000209225. *A. queenslandica* orthologies were delineated using OrthoMCL[35], and those of *X. tropicalis* and *N. vectensis* were retrieved from Biomart[36] which contained the annotations on the noted versions. We included in the analysis genes whose maximum expression was greater than the data-set-specific threshold; this was computed as the median average expression across all genes. Expression profiles passing this threshold were each normalized to their own maximum expression. A Kolmogorov–Smirnov test was used to test for significantly different temporal dynamics between endoderm and ectoderm expression. For this analysis the timing of expression for each gene was computed as the stage at which half of the sum expression had occurred.

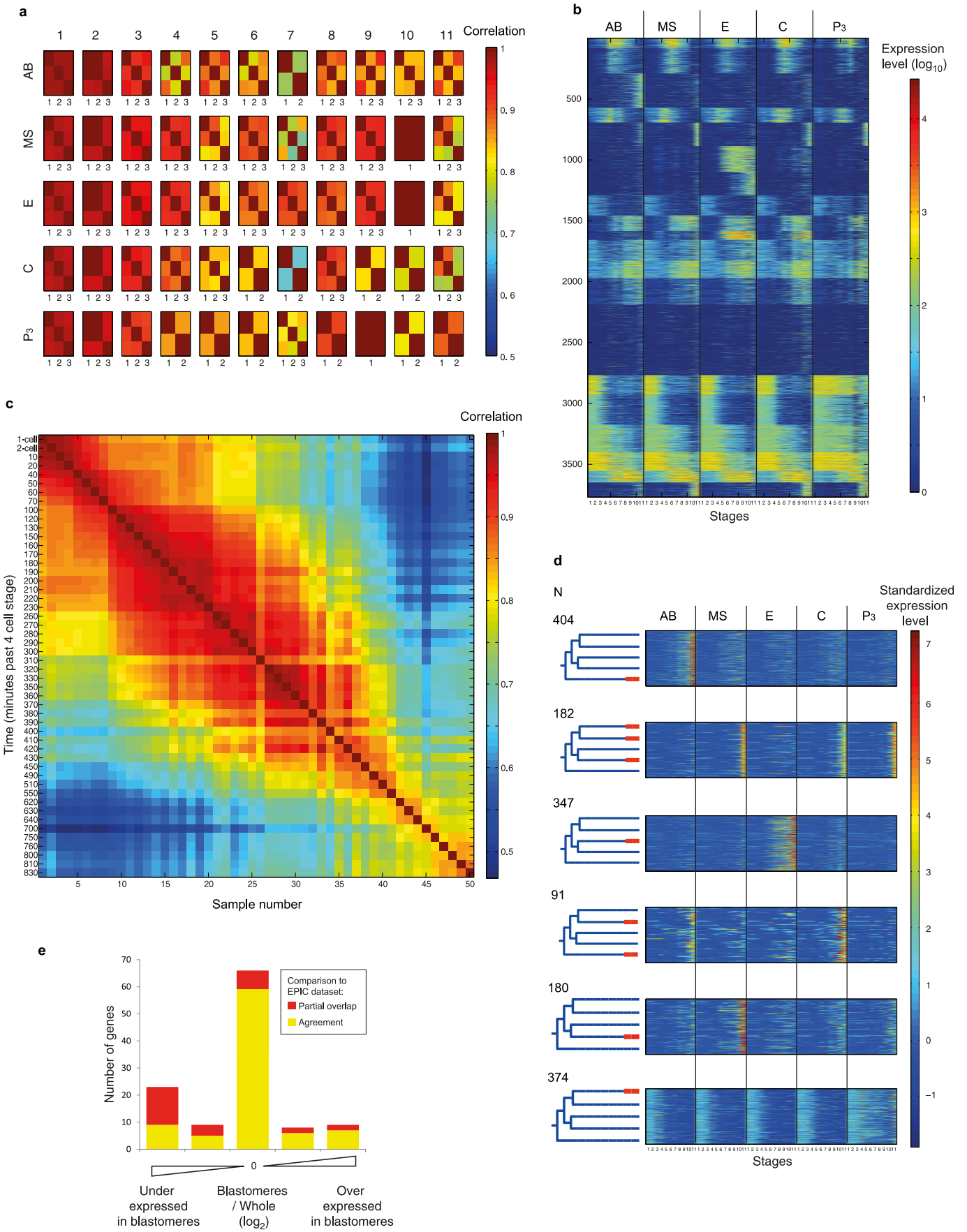
Functional categories analysis. COG[37] functional category annotations were retrieved from WormMart[38]. For simplicity, annotations of 'general function prediction only' and 'function unknown' were ignored, as well as those categories capturing fewer than 3% of the genes. Enrichments were computed using the hypergeometric distribution.

31. Goldstein, B. Induction of gut in *Caenorhabditis elegans* embryos. *Nature* **357,** 255–257 (1992).
32. Baker, S. C. *et al.* The External RNA Controls Consortium: a progress report. *Nature Methods* **2,** 731–734 (2005).
33. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
34. Pryszcz, L. P., Huerta-Cepas, J. & Gabaldon, T. MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. *Nucleic Acids Res.* **39,** e32 (2011).
35. Fischer, S. *et al.* in *Current Protocols in Bioinformatics* Ch. 6, Unit 6.12, 11–19 (2011).
36. Guberman, J. M. *et al.* BioMart Central Portal: an open database network for the biological community. *Database* **2011,** bar041 (2011).
37. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science* **278,** 631–637 (1997).
38. Schwarz, E. M. *et al.* WormBase: better software, richer content. *Nucleic Acids Res.* **34,** D475–D478 (2006).
39. Murray, J. I. *et al.* Automated analysis of embryonic gene expression with cellular resolution in *C. elegans*. *Nature Methods* **5,** 703–709 (2008).
40. Cowan, A. E. & McIntosh, J. R. Mapping the distribution of differentiation potential for intestine, muscle, and hypodermis during early development in *Caenorhabditis elegans*. *Cell* **41,** 923–932 (1985).

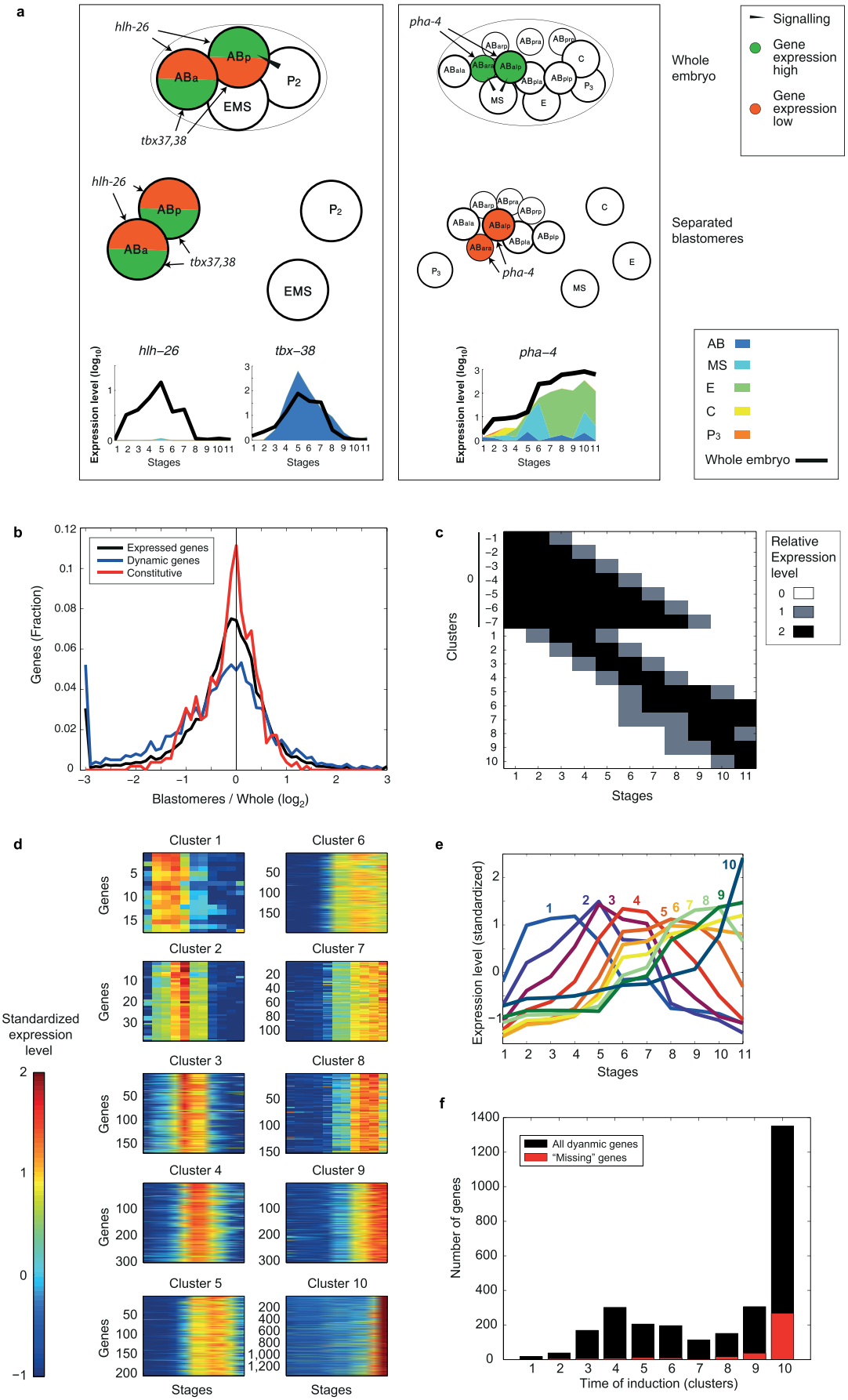
41. Good, K. *et al.* The T-box transcription factors TBX-37 and TBX-38 link GLP-1/Notch signaling to mesoderm induction in *C. elegans* embryos. *Development* **131,** 1967–1978 (2004).
42. Goldstein, B. An analysis of the response to gut induction in the *C. elegans* embryo. *Development* **121,** 1227–1236 (1995).
43. Laufer, J. S., Bazzicalupo, P. & Wood, W. B. Segregation of developmental potential in early embryos of *Caenorhabditis elegans*. *Cell* **19,** 569–577 (1980).
44. Goldstein, B. Establishment of gut fate in the E lineage of *C. elegans*: the roles of lineage-dependent mechanisms and cell interactions. *Development* **118,** 1267–1277 (1993).
45. Fox, R. M. *et al.* The embryonic muscle transcriptome of *Caenorhabditis elegans*. *Genome Biol.* **8,** R188 (2007).
46. McGhee, J. D. *et al.* ELT-2 is the predominant transcription factor controlling differentiation and function of the *C. elegans* intestine, from embryo to adult. *Dev. Biol.* **327,** 551–565 (2009).



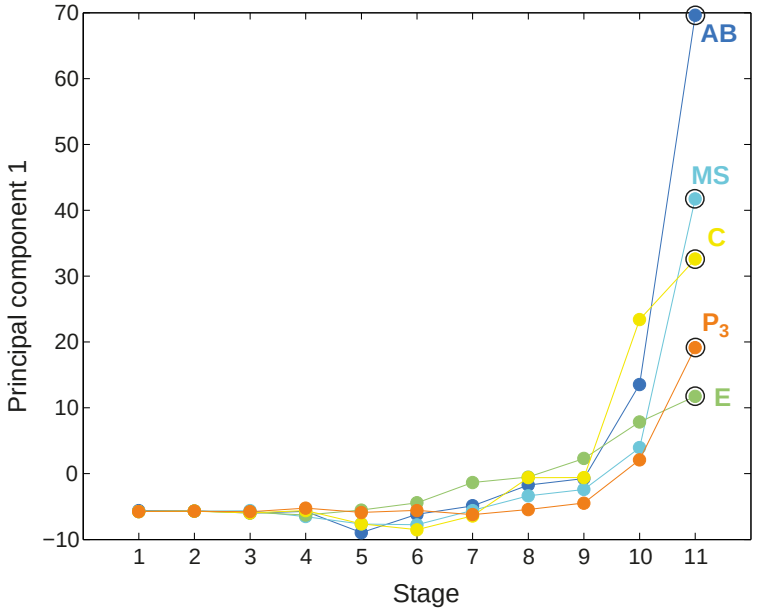
Extended Data Figure 1 | *In vitro* culturing of the *C. elegans* embryonic founder blastomeres. The cells are separated as shown in the left schematic and then cultured in embryonic growth medium[11] as shown in the micrographs on the right. The numbers indicate the stages at which the cells were collected for transcriptome analysis. Six of the 11 stages are shown in the micrographs.



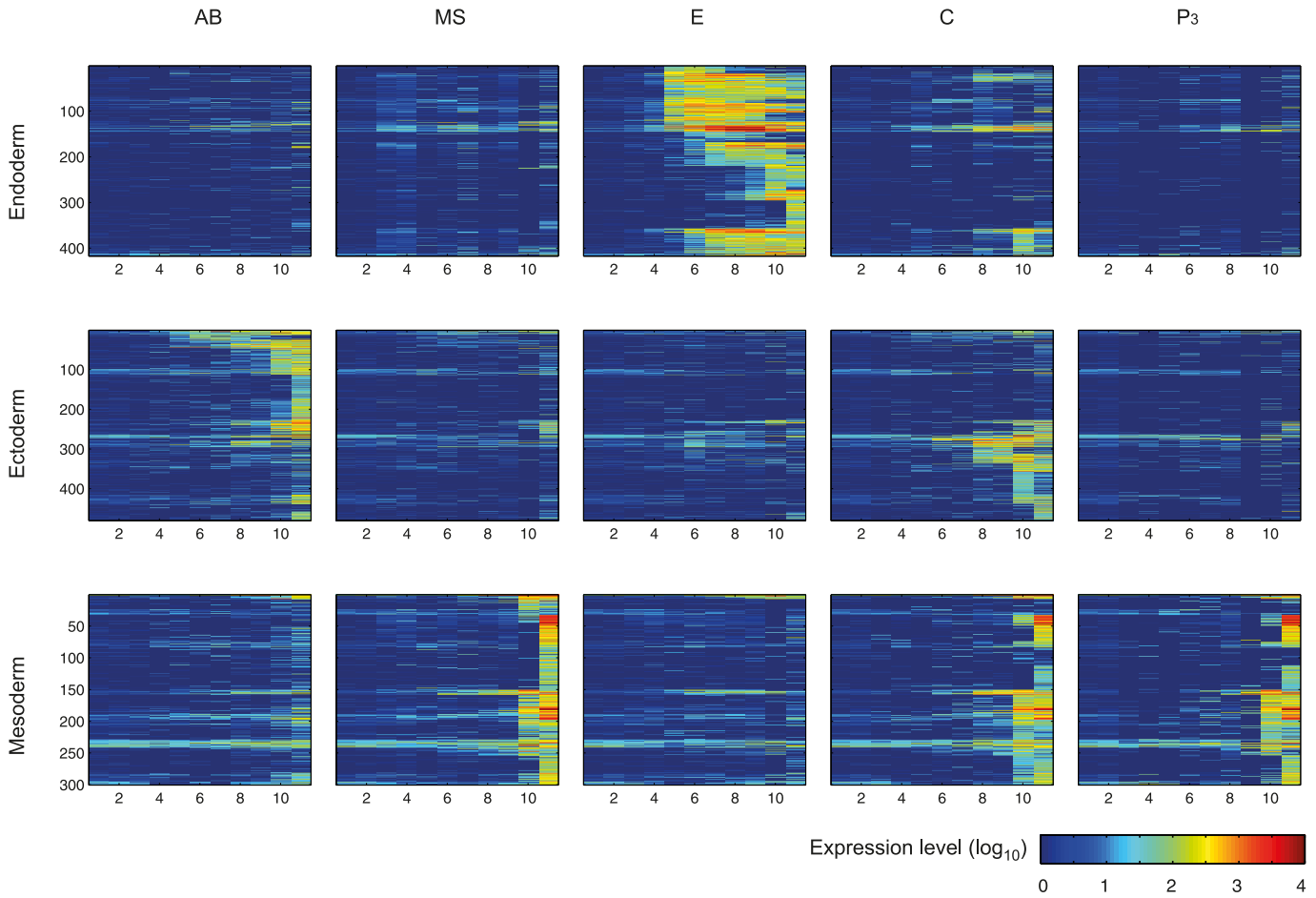
Extended Data Figure 2 | A transcriptomic survey of *C. elegans* embryonic founder cell lineages. **a**, Replicates of the embryonic blastomere time courses. The heat maps show the correlations among the replicates for each blastomere lineage at each of the eleven examined stages. For three blastomere stages there were no replicates. The median correlation coefficient is 0.9. Samples were collected in triplicates. Only samples with at least 750,000 reads were used, which has been previously shown to be of sufficient sequencing depth for CEL-seq[12]. Supplementary Table 3 provides the sequencing statistics for each sample. **b**, Expression profiles of the 3,910 dynamic genes across the blastomere lineage time courses. See Methods for definition of dynamic genes. **c**, Correlation coefficients between samples of the whole-embryo time course. Each of the 50 samples comprises a single embryo, collected at the indicated minutes past the four-cell stage. Again, only samples with at least 750,000 reads were used and Supplementary Table 3 provides the sequencing statistics for each sample. **d**, The expression profiles of the 1,664 genes with differentiated expression analysed in Fig. 1c. Each profile was 'standardized' by subtracting its mean and dividing by its standard deviation. **e**, Comparison of the blastomere time courses to the EPIC data set[15]. For 115 genes, we could compare gene expression to previously published embryonic expression profiles generated by microscopic lineaging until the ~300-cell stage[15,39]. Of these, 75% of our profiles had consistent localized expression (Supplementary Table 1). Of those, 54% matched completely, and 21% of the genes expressed in all of the lineages in our data set had some missing expression in the EPIC data set because the lineaging was not performed until the end of the developmental process. The remaining genes have some overlap in expression. Such differences in expression could be caused by the transgene in the EPIC data set not recapitulating the profile of the endogenous gene, or missing signals between cells in the blastomere data set, as is seen from the whole-embryo/blastomere expression level ratio (see Supplementary Table 1, ratios defined as equal, slightly higher/lower or much higher/lower). Expression profile compared with the EPIC data set deviates more when expression in the blastomeres is low compared with the whole embryo, but the blastomere data set has the advantage that all genes are assayed simultaneously, no transgenes are used, maternal transcripts are seen and downregulation of genes is observable.



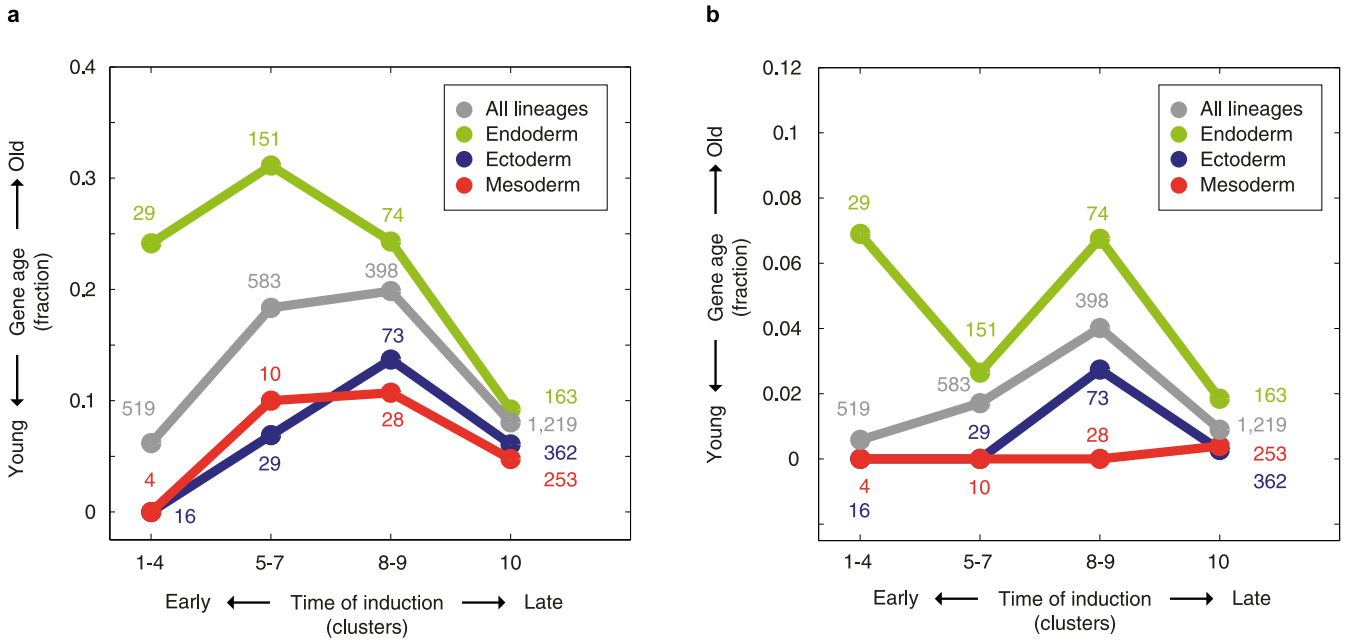
Extended Data Figure 3 | Lineage-restricted gene expression identifies genes dependent upon coherence of the lineages and tissue specificity. **a**, Expression profiles of genes involved in pharynx specification. The left and right panels correspond to the two Notch signalling events. The top and bottom images correspond to the expected regulatory patterns in the whole embryo and isolated blastomeres, respectively. The *tbx-37* gene is not shown since it is identical to *tbx-38* in expression profile. **b**, Comparison of the overall sum of expression between the two time courses, plotted on a $\log_2$ scale (black). Genes 'missing' in the separated lineage time course were manually added to the graph at −3. The additional plots indicate the same measure for dynamically expressed genes (blue) and constitutive genes (red). **c**, Idealized expression profiles used to identify gene expression clusters. **d**, The gene expression profiles for the temporally restricted gene expression profiles. Each profile was 'standardized' by subtracting its mean and dividing by its standard deviation. **e**, Average expression profiles of ten clusters of dynamically expressed genes determined on the basis of the whole-embryo expression data (see Methods). **f**, The number of dynamic genes in each temporal period. In each group, the genes not expressed in the lineage time course (**b**) are marked in red.



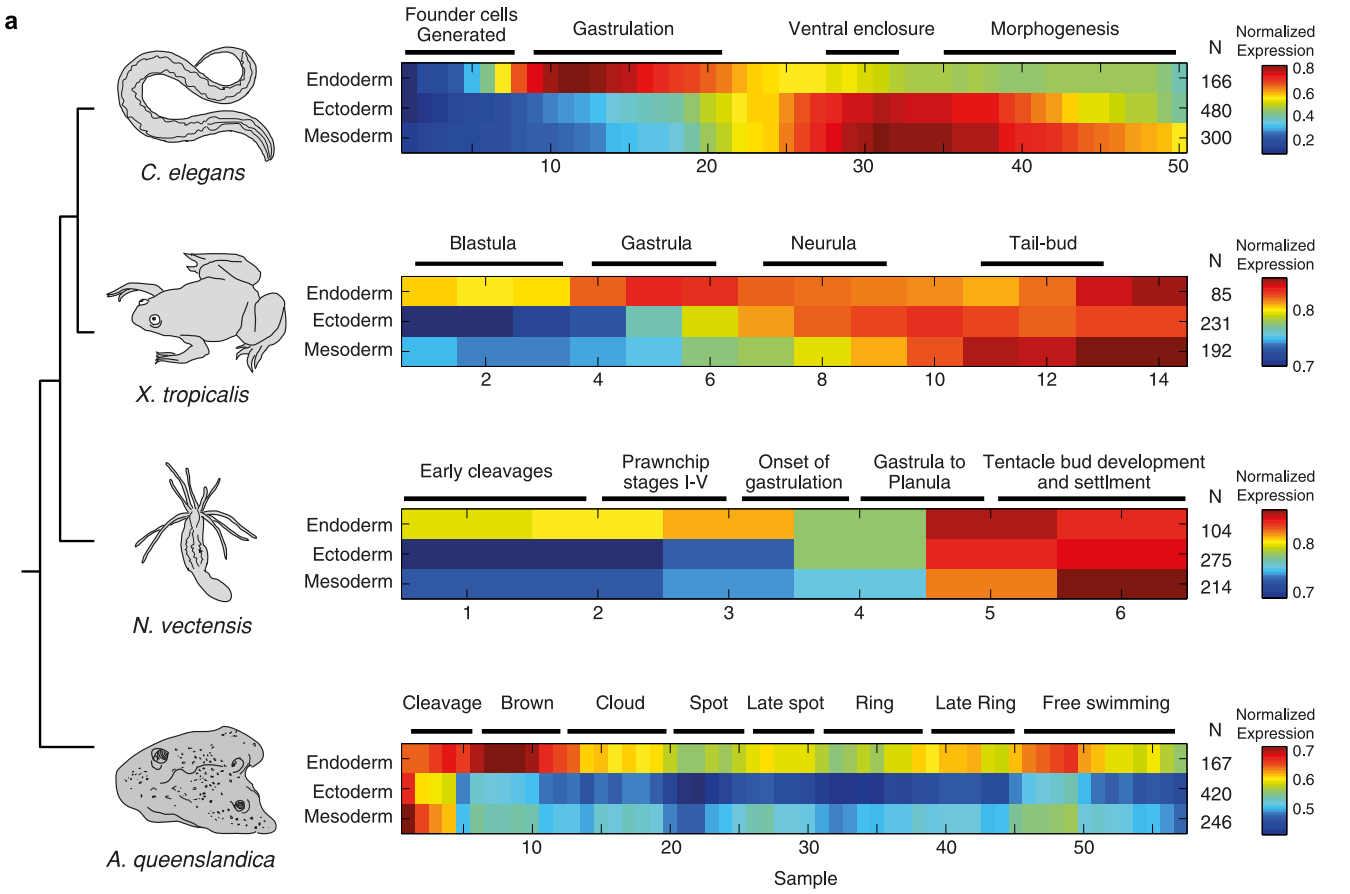
Extended Data Figure 4 | The first principal component correlates with developmental time. Principal component analysis as described in Fig. 2a. Colour codes are the same as in Fig. 1. PC1, PC2 and PC3 capture 18%, 12% and 11%, respectively, of the variation in the expression, in the 1,320 dynamically expressed genes with no expression in the first stage (to exclude genes with maternal expression).

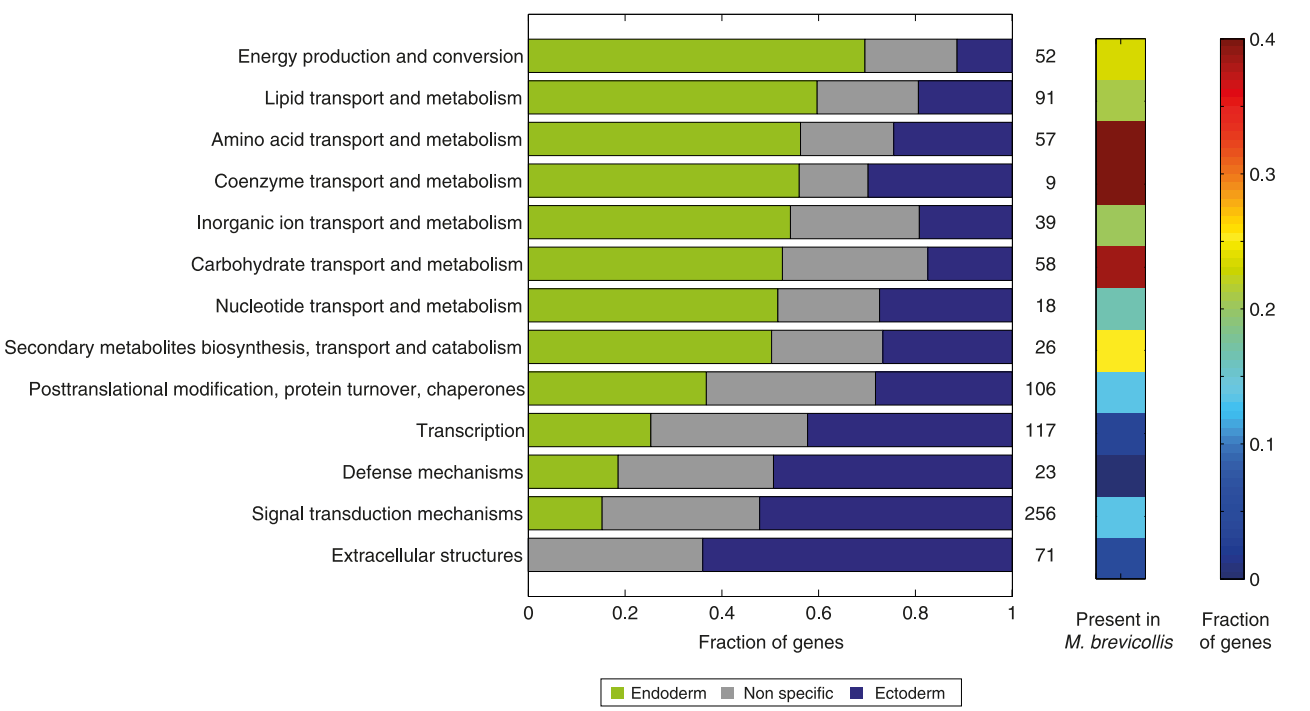


Extended Data Figure 5 | Germ-layer-specific expression. Expression profiles of the germ-layer-specific genes in each of the lineages. The *x* and *y* axes are the 11 examined temporal stages and individual genes, respectively. Germ-layer-specific genes were identified by hierarchical clustering based upon correlation among dynamically expressed genes (see Methods).



Extended Data Figure 6 | Robustness of gene age analysis. **a**, Same format as Fig. 4a but with the definition of old genes as those present in at least 25% of the examined eukaryotes (see Methods) that are not ophisthokonts. **b**, Same as Fig. 4a with a definition of 'old' as those present in 25% of the examined organisms that are not eukaryotes (Eubacteria and Archaea).





Extended Data Figure 7 | Truncated endoderm gene set control. To exclude the possibility that general genes were included as 'endoderm-specific' because the endoderm program is induced earlier, we excluded temporal clusters 8, 9 and 10 from the endoderm genes and repeated the relevant analyses. We found that there was no marked change in the results. The results are shown in the same format as Figs 3 and 4b, c.

Extended Data Table 1 | The fates of the progeny of each blastomere *in vivo* and in isolated cultured blastomeres

		Fates in whole embryo [10]	Expected *in vitro*	References
AB		Neurons	Unknown	
		Epidermis	Yes	40
		Pharynx	No	41
		1 muscle cell	Unknown	
MS		Muscle	Yes	42
		Pharynx	Yes	42
E		Endoderm	Yes	43,44
C		Muscle	Yes	40
		Epidermis	Yes	40
P3	D	Muscle	Yes	40
	P4	Germ line	Unknown	

Data are from refs 40–44.

Extended Data Table 2 | Description of the developmental stages queried in this study

Stage number	Stage name	Description	Time*
1	2-cell	2-cell embryo	0
2	4-cell	4-cell embryo	20
3	E	After division of EMS to E and MS	40
4	2E	After division of E to Ea and Ep	60
5	2E+	After division of MSa and MSp to MSaa, MSap, MSpa and MSpp	90
6	4E	After division of Ea and Ep to Eal, Ear, Epl and Epr	110
7	4E+	60 minutes after division of Ea and Ep to Eal, Ear, Epl and Epr	140
8	8E	After division of Eal, Ear, Epl and Epr to Eala, Ealp, Eara, Earp, Epla, Eplp, Epra and Eprp	180
9	8E+	90 minutes after division of Eal, Ear, Epl and Epr to Eala, Ealp, Eara, Earp, Epla, Eplp, Epra and Epr	na
10	8E++	180 minutes after division of Eal, Ear, Epl and Epr to Eala, Ealp, Eara, Earp, Epla, Eplp, Epra and Epr	na
11	o.n.	After an over-night incubation – more than 8 E cells are visible.	na

* Timing of the stage in the Sulston lineage[10]. Timing is indicated as minutes from the 2-cell stage.

Extended Data Table 3 | Tissue-specific gene sets

Tissue	Gene sets
Neuronal	Genes with the following GO terms: GO:0001764 neuron migration GO:0004983 neuropeptide Y receptor activity GO:0005328 neurotransmitter:sodium symporter activity GO:0006836 neurotransmitter transport GO:0007218 neuropeptide signaling pathway GO:0007268 synaptic transmission GO:0007411 axon guidance GO:0008021 synaptic vesicle GO:0030424 axon GO:0030425 dendrite GO:0030594 neurotransmitter receptor activity GO:0043005 neuron projection GO:0045202 synapse GO:0045211 postsynaptic membrane GO:0048489 synaptic vesicle transport GO:0048666 neuron development
Muscle	Genes identified by Fox et al.[45]
Endoderm	Genes identified by McGhee et al.[46]
Epidermis	Genes with the following GO term: GO:0018996 molting cycle, collagen and cuticulin-based cuticle
Pharynx	Genes with the following GO term: GO:0007631 feeding behavior
Germline	Genes with the following GO terms: GO:0051729 germline cell cycle switching, mitotic to meiotic cell cycle GO:0048477 oogenesis GO:0045132 meiotic chromosome segregation GO:0043186 P granule GO:0007276 gamete generation GO:0007281 germ cell development GO:0007126 meiosis GO:0001556 oocyte maturation GO:0000003 reproduction

Data for muscle and endoderm are from refs 45 and 46, respectively.

LETTER

doi:10.1038/nature14135

Large-scale discovery of novel genetic causes of developmental disorders

The Deciphering Developmental Disorders Study*

Despite three decades of successful, predominantly phenotype-driven discovery of the genetic causes of monogenic disorders[1], up to half of children with severe developmental disorders of probable genetic origin remain without a genetic diagnosis. Particularly challenging are those disorders rare enough to have eluded recognition as a discrete clinical entity, those with highly variable clinical manifestations, and those that are difficult to distinguish from other, very similar, disorders. Here we demonstrate the power of using an unbiased genotype-driven approach[2] to identify subsets of patients with similar disorders. By studying 1,133 children with severe, undiagnosed developmental disorders, and their parents, using a combination of exome sequencing[3–11] and array-based detection of chromosomal rearrangements, we discovered 12 novel genes associated with developmental disorders. These newly implicated genes increase by 10% (from 28% to 31%) the proportion of children that could be diagnosed. Clustering of missense mutations in six of these newly implicated genes suggests that normal development is being perturbed by an activating or dominant-negative mechanism. Our findings demonstrate the value of adopting a comprehensive strategy, both genome-wide and nationwide, to elucidate the underlying causes of rare genetic disorders.

We established a network to recruit 1,133 children (median age 5.5 years, Extended Data Fig. 1a) with diverse, severe undiagnosed developmental disorders, through all 24 regional genetics services of the UK National Health Service and Republic of Ireland. Among the most commonly observed phenotypes (Extended Data Fig. 1b and Supplementary Table 1) were intellectual disability or developmental delay (87% of children), abnormalities revealed by cranial MRI (30%), seizures (24%), and congenital heart defects (11%). These children are predominantly (~90%) of northwest European ancestry (Extended Data Fig. 1c), with 47 pairs of parents (4.1%) exhibiting kinship equivalent to, or in excess of, second cousins (Extended Data Fig. 1d and Supplementary Information). In most families (849 of 1,101) the child was the only affected family member, but 111 children had one or more parents with a similar developmental disorder, and 124 had a similarly affected sibling (Supplementary Information). Prior clinical genetic testing would have already diagnosed many children with easily recognized syndromes, or large pathogenic deletions and duplications, enriching this research cohort for less distinct syndromes and novel genetic disorders.

We sequenced the exomes of 1,133 children with developmental disorders and their parents, from 1,101 families, representing 1,071 unrelated children and 30 sibships. We also performed exome-focused array comparative genomic hybridization (exome-aCGH) on the children ($n = 1{,}009$) and UK controls ($n = 1{,}013$), and genome-wide genotyping on the trios ($n = 1{,}006$) to identify deletions, duplications, uniparental disomy and mosaic large chromosome rearrangements. From our exome sequencing and exome-aCGH data, we detected an average of 19,811 coding or splicing single nucleotide variants (SNVs), 491 coding or splicing insertions and deletions (indels) and 148 copy number variants (CNVs) per child (Supplementary Information). From analyses of the genotyping array data[12] we identified six children with uniparental disomy and five children with mosaic large chromosomal rearrangements (Supplementary Information). The SNVs, indels and CNVs were analysed jointly in the following analyses, allowing, for example, the identification of compound heterozygous CNVs and SNVs affecting the same gene.

We discovered 1,618 *de novo* variants (1,417 SNVs, 114 indels and 87 CNVs) in coding and non-coding regions (Supplementary Tables 2 and 3), of which 1,596 (98.6%) were validated using a second, independent assay, and the remainder were validated clinically. This represents an average of 1.12 *de novo* SNVs and 0.09 *de novo* indels in coding or splicing regions per child, which is within the range of similar studies[3–11]. The distribution of *de novo* SNVs and indels per child closely approximated the Poisson distribution expected for random mutational events (Extended Data Fig. 2).

We classified 28% ($n = 317$) of children with probable pathogenic variants (Supplementary Table 4 and ref. 13) in 1,129 robustly implicated developmental disorder genes (published before November 2013), or with pathogenic deletions or duplications. Most of these diagnoses involved *de novo* SNVs, indels or CNVs (Table 1). Females had a significantly higher diagnostic yield of autosomal *de novo* mutations than males ($P = 0.01$, Fisher's exact test). Among the single-gene diagnoses, most genes linked to developmental disorders (95 out of 148) were only observed once, although eight (*ARID1B*, *SATB2*, *SYNGAP1*, *ANKRD11*, *SCN1A*, *DYRK1A*, *STXBP1*, *MED13L*) each accounted for 0.5–1% of children in our cohort (Extended Data Fig. 3). For seventeen of these children we identified two different genes with pathogenic variants, resulting in a composite clinical phenotype.

Analyses that assess the enrichment in patients of a particular class of variation, so-called 'burden analyses', both highlight classes of variants for detailed analysis and enable estimation of the proportion of a particular class of variant that is likely to be pathogenic. We observed a significant ($P = 0.0004$) burden of 87 *de novo* CNVs in the 1,133 children with developmental disorders compared to 12 in 416 controls (Scottish Family Health Study[14]) despite most children (77%) having previously had clinical microarray testing (Extended Data Fig. 4).

We used gene-specific mutation rates that account for gene length and sequence context[15] to assess the burden of different classes of *de novo* SNVs and indels (Supplementary Information). We observed no significant excess of any functional class of *de novo* SNVs or indels in

Table 1 | Breakdown of diagnoses by mode and by sex

	Female (%)	Male (%)	Total (%)
Undiagnosed	383 (69.6)	433 (74.3)	816 (72.0)
Diagnosed	167 (30.4)	150 (25.7)	317 (28.0)
De novo mutation	124 (22.5)	80 (13.7)	204 (18.0)
Chr X*	24 (4.4)	5 (0.9)	28 (2.6)
Autosomal*	100 (18.2)	75 (12.9)	176 (15.5)
Autosomal dominant†	9 (1.6)	11 (1.9)	20 (1.8)
Autosomal recessive	20 (3.6)	26 (4.5)	46 (4.1)
X-linked inherited	1 (0.2)	19 (3.3)	20 (1.8)
UPD/mosaicism	4 (0.7)	6 (1.0)	10 (0.9)
Composite	9 (1.6)	8 (1.4)	17 (1.5)
Total	550	583	1,133

UPD, uniparental disomy.
* Chromosome X (Chr X) and autosomal values are subsets of '*De novo* mutation'.
† Inherited from a parent.

*Lists of participants and their affiliations appear at the end of the paper.

autosomal-recessive developmental-disorder-linked genes (Extended Data Fig. 5), suggesting that few of these mutations are causally implicated. By contrast, we observed a highly significant excess of all 'functional' classes (coding and splice site variants excepting synonymous changes) of *de novo* SNVs and indels in the dominant and X-linked developmental-disorder-linked genes (Extended Data Fig. 5) within which *de novo* mutations can be sufficient to cause disease. Not all protein-altering mutations in known dominant and X-linked developmental disorder genes will be pathogenic, and these burden analyses inform estimates of positive predictive values for different classes of mutations. The remaining genes (that is, those not linked to developmental disorder) in the genome also exhibit a more modest, but significant, excess of functional, but not silent, *de novo* SNVs and indels (Extended Data Fig. 5).

We observed 96 genes with recurrent, functional mutations (Fig. 1a), a highly significant excess compared to the expected number derived from simulations (median = 55; Supplementary Information). This enrichment is even more pronounced (observed, 29; expected, 3) for recurrent loss-of-function mutations (Fig. 1b). Among undiagnosed children, we observed an excess of 22 genes (observed: 45, expected: 23) with recurrent functional mutations (Fig. 1a) and an excess of 8 genes (observed, 9; expected, 1) with recurrent loss-of-function mutations (Fig. 1b), implying that an appreciable fraction of these recurrently mutated genes are novel developmental-disorder-linked genes.

To identify individual genes enriched for damaging *de novo* mutations (Supplementary Information), we tested for a gene-specific overabundance of either *de novo* loss-of-function mutations or clustered functional *de novo* mutations in 1,130 children (excluding one twin from each of three identical twin pairs). To increase power to detect genes associated with developmental disorder, we also meta-analysed our data with published *de novo* mutations from 2,347 developmental disorder trios with intellectual disability[4,9], epileptic encephalopathy[3], autism[6–8,10], schizophrenia[5], or congenital heart defects[11] (the 'meta-DD' data set). These analyses (Fig. 2) successfully re-discovered 20 known genes linked to developmental disorder at genome-wide significance ($P < 1.31 \times 10^{-6}$, a Bonferroni P value of 0.05 corrected for 38,504 tests (Supplementary Information)). Thus, despite the broad phenotypic ascertainment in these data sets, we can robustly detect developmental-disorder-linked genes solely on statistical grounds.

To increase our power to detect novel genes linked to developmental disorder, we repeated the gene-specific analysis described above excluding the 317 individuals with a known cause of their developmental disorder. In this analysis the statistical genetic evidence was integrated with phenotypic similarity of patients, available data on model organisms and functional plausibility. We identified twelve novel disease genes with compelling evidence for pathogenicity (Table 2), nine of which

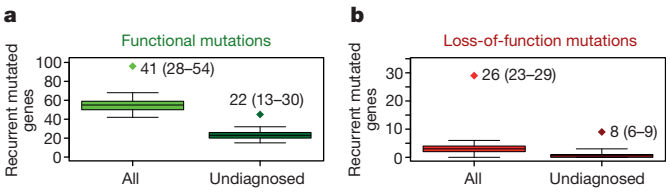


Figure 1 | Excess of recurrently mutated genes. Each panel shows the observed number of recurrently mutated genes (diamond) and the distribution of the number of recurrently mutated genes in 10,000 simulations (box indicates interquartile range, whiskers indicates 95% confidence interval) under a model of no gene-specific enrichment of mutations. **a**, All protein-altering mutations in all DDD children and undiagnosed DDD children. **b**, All loss-of-function mutations in all DDD children and undiagnosed DDD children. Each diamond is annotated with the median excess of recurrently mutated genes, with 95% confidence intervals in brackets. P value of observed excess is <0.0001 for all four tests. No statistical methods were used to predetermine sample size.

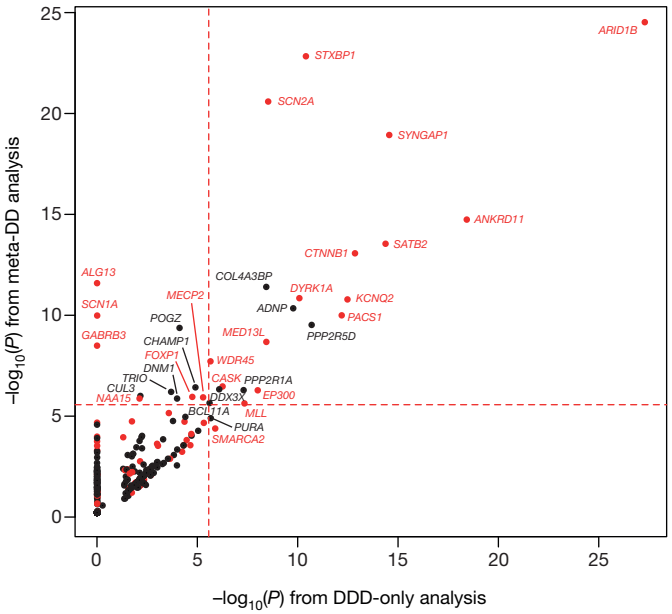


Figure 2 | Gene-specific significance of enrichment for *de novo* mutations. The $-\log_{10}(P)$ value of testing for mutation enrichment is plotted only for each gene with at least one mutation in DDD children. On the x axis is the P value of the most significant test in the DDD data set; on the y axis is the minimal P value from the significance testing in the meta-analysis data set. Red indicates genes already known to be associated with developmental disorders (in DDG2P). Only genes with a P value of less than 0.05/18,272 (red lines) (where 18,272 is the number of genes tested) are labelled.

exceeded the genome-wide significance threshold of 1.36×10^{-6} (Supplementary Information), with the remaining three genes (*PCGF2*, *DNM1* and *TRIO*) just below this significance threshold. The two children with identical Pro65Leu mutations in *PCGF2*, which encodes a component of a Polycomb transcriptional repressor complex, share a strikingly similar facial appearance representing a novel and distinct dysmorphic syndrome. *DNM1* was previously identified as a candidate gene for epileptic encephalopathy[3]. Two of the three children that we identified with *DNM1* mutations also had seizures, and a heterozygous mouse mutant manifests seizures[16]. In addition to two *de novo* missense SNVs in *TRIO*, we identified an intragenic *de novo* 82-kilobase (kb) deletion of 16 exons. For several of these novel developmental-disorder-linked genes, the meta-DD analysis increased the significance of enrichment. For example, a total of five *de novo* loss-of-function variants in *POGZ* were identified, two from our cohort, two from recent autism studies[6,7] and one from a recent schizophrenia study[5]. We also identified six genes with suggestive statistical evidence of being novel genes associated with developmental disorder, defined as having a P value for mutation enrichment less than 1×10^{-4} and being plausible from a functional perspective (Extended Data Table 1). We anticipate that most of these genes will eventually accrue sufficient evidence to meet the stringent criteria we defined above for declaring a novel developmental-disorder-linked gene.

Notably, we observed identical missense mutations in unrelated, phenotypically similar patients for four of these novel developmental-disorder-linked genes (*PCGF2*, *COL4A3BP*, *PPP2R1A* and *PPP2R5D*), and for a fifth gene, *BCL11A*, we identified highly significant clustering of non-identical missense mutations (Fig. 3). We hypothesize that the mutations in some of these genes may be operating by either dominant-negative or activating mechanisms. This hypothesis is supported by previous functional evidence for several of the mutated amino acids. The three identical Ser132Leu mutations in *COL4A3BP*, which encodes an intracellular transporter of ceramide, remove a serine that when phosphorylated downregulates transporter activity from the ER to the

Table 2 | Novel genes with compelling evidence for a role in developmental disorder

Evidence	Gene	*De novo* DDD (missense, LOF)	*De novo* meta (missense, LOF)	*P* value	Test	Mutation clustering	Predicted haploinsufficiency (%)
De novo enrichment	*COL4A3BP*	3 (3,0)	5 (5,0)	4.10×10^{-12}	Meta	Yes	14.7
	PPP2R5D	4 (4,0)	5 (5,0)	6.01×10^{-12}	DDD	Yes	19.7
	ADNP	4 (0,4)	5 (0,5)	4.59×10^{-11}	Meta	No	9.8
	POGZ	2 (0,2)	5 (0,5)	4.31×10^{-10}	Meta	No	30.0
	PPP2R1A	3 (3,0)	3 (3,0)	2.03×10^{-8}	DDD	Yes	23.5
	DDX3X	4 (3,1)	5 (3,2)	2.26×10^{-7}	DDD	No	12.7
	CHAMP1	2 (0,2)	3 (0,3)	4.58×10^{-7}	Meta	No	52.9
	BCL11A	3 (3,0)	4 (3,1)	1.03×10^{-6}	DDD	Yes	0.6
	PURA	3 (1,2)	3 (1,2)	1.14×10^{-6}	DDD	No	9.4
De novo enrichment plus additional evidence	*DNM1*	3 (3,0)	5 (5,0)	1.43×10^{-6}	Meta	No	13.5
	TRIO	2 (2,0)	7 (7,0)	5.16×10^{-6}	Meta	Yes	25.7
	PCGF2	2 (2,0)	2 (2,0)	1.08×10^{-5}	DDD	Yes	37.7

The table summarizes the 12 genes with compelling evidence to be novel developmental-disorder-linked genes. The number of unrelated patients with independent functional or loss-of-function (LOF) mutations in the Deciphering Developmental Disorders (DDD) cohort or the wider meta-analysis (meta) data set including DDD patients is listed. The *P* value reported is the minimum *P* value from the testing of the DDD data set and the meta-analysis data set. The data set that gave this minimal *P* value is also reported. Mutations are considered to be clustered if the *P* value of clustering of functional SNVs is less than 0.01. Predicted haploinsufficiency is reported as a percentile of all genes in the genome, with ~0% being highly likely to be haploinsufficient and 100% very unlikely to be haploinsufficient, based on the prediction score described in ref. 26 updated to enable predictions for a higher fraction of genes in the genome. During submission, a paper was published describing a novel developmental disorder caused by mutations in *ADNP* (ref. 27).

Golgi[17], presumably resulting in intra-cellular imbalances in ceramide and its downstream metabolic pathways. The two mutated amino acids (Arg182Trp and Pro179Leu) in *PPP2R1A*, which encodes the scaffolding A subunit of the protein phosphatase 2 complex, have been previously identified as sites of driver mutations in endometrial and ovarian cancer[18]. It has previously been shown that mutating either of these two residues results in impaired binding of B subunits of the complex[18]. Intriguingly, *PPP2R5D* encodes one of the possible B subunits of the same protein phosphatase 2 complex, suggesting that the clustered missense mutations (Pro201Arg and Glu198Lys) in this gene may similarly perturb interactions between subunits of this complex. Further functional studies will be required to confirm this hypothesis.

We assessed transmission biases of potentially pathogenic inherited SNVs in our probands (Supplementary Information) and observed a genome-wide trend ($P = 0.015$) towards over-transmission to probands of very rare (minor allele frequency (MAF) <0.0005%) loss-of-function variants, but not damaging missense variants. We also observed a 1.8-fold enrichment ($P = 0.04$) of rare (MAF <5%) biallelic loss-of-function variants (Supplementary Table 5) among probands without a likely dominant cause of their disorder, compared to those with either a diagnostic *de novo* mutation or an affected parent. Again we saw no enrichment in biallelic damaging missense variants (Extended Data Table 2), consistent with a similar observation in children with autism[19]. These observations suggest that although inherited loss-of-function variants (both monoallelic and biallelic) are probably contributing to developmental disorder in our patients, much larger sample sizes will be required to pinpoint specific developmental-disorder-linked genes in this way.

To direct future, detailed functional experiments on the developmental role of a subset of candidate genes from this study we used two approaches. First, knockdown-induced phenotypes were recorded in early zebrafish development. Second, we performed a systematic review of perturbed gene function in human, mouse, *Xenopus*, zebrafish and *Drosophila*. In both approaches the animal phenotypes were compared to those seen in individuals in our cohort.

We undertook an antisense-based loss-of-function screen in zebrafish to assess 32 candidate developmental-disorder-linked genes with *de novo* loss-of-function, *de novo* missense or biallelic loss-of-function variants from exome sequencing (Supplementary Information and Supplementary Table 6). These candidate genes corresponded to 39 zebrafish orthologues. Knockdowns of these zebrafish genes were repeated at least twice and all morpholinos were co-injected with *tp53* morpholino to eliminate off-target toxicity. Successful knockdown of the targeted messenger RNA could be confirmed using polymerase chain reaction with reverse transcription (RT–PCR) for 82.4% of genes (28 out of 34), and 9 out of 11 (82%) of genes that were tested gave an equivalent phenotype when knocked down by a second, independent morpholino. Knockdown of at least one or a pair of zebrafish orthologues of 65.6% of candidate developmental-disorder-linked genes (21 out of 32) resulted in perturbed embryonic and larval development (Fig. 4, Extended Data Table 3, Supplementary Data and Supplementary Table 7). Large-scale mutagenesis[20] and morpholino[21] studies suggest that knockout or knockdown of 6–12% of genes give developmental phenotypes, suggesting at least a fivefold enrichment of developmentally non-redundant genes among the 32 selected for modelling. We then compared the phenotypes of the zebrafish morphants to those of the patients with *de novo* mutations or biallelic loss-of-function variants in the orthologous genes (Extended Data Table 3). Eleven out of twenty-one (52.4%) of the genes were categorized as strong candidates based on phenotypic similarity (Fig. 4a). Seven out of eleven were potential microcephaly genes, the knockdown of which in zebrafish gives significant reductions in both head measurements and neural tissue (Fig. 4b and Supplementary Information). Six out of twenty-one (28.6%) genes resulted in severe morphant

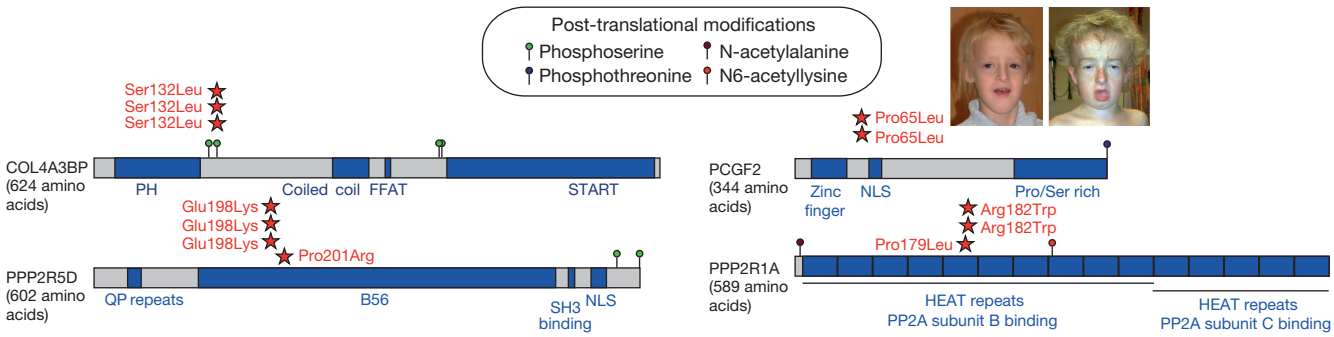


Figure 3 | Four novel genes with clustered mutations. The domains (blue), post-translational modifications, and mutation locations (red stars) are shown for four proteins with highly clustered *de novo* mutations in unrelated children with severe, undiagnosed developmental disorders. For the protein PCGF2, where all observed mutations are identical, photos are shown to highlight the facial similarities of patients carrying the same mutation.

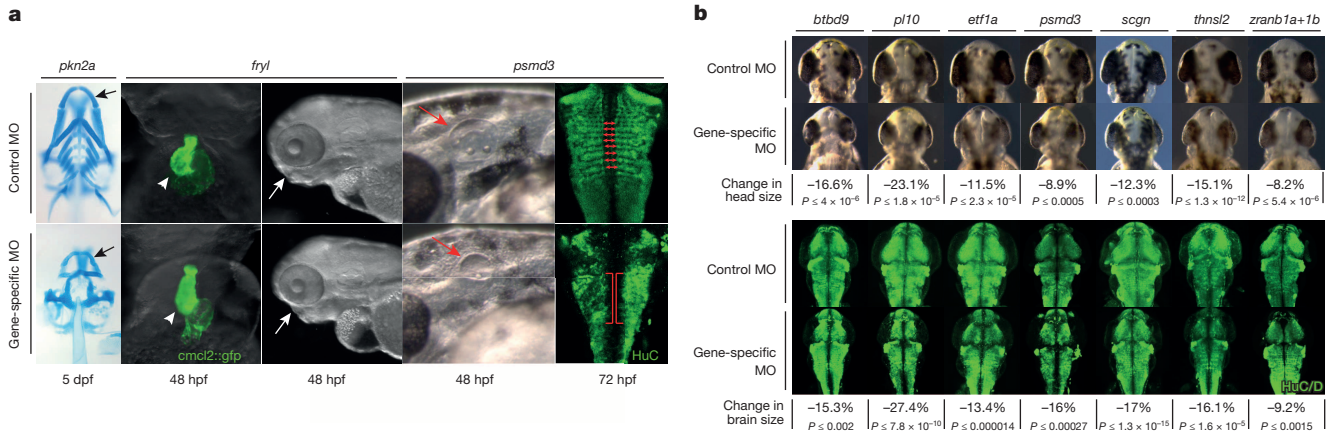


Figure 4 | Candidate gene loss-of-function modelling in zebrafish reveals enrichment for developmentally important proteins. **a**, Examples of developmental phenotypes: knockdown of *pkn2a* results in reduced cartilaginous jaw structures (black arrows); knockdown of *fryl* results in cardiac and craniofacial defects (white arrowheads and arrows, respectively); while knockdown of *psmd3* results in smaller ear primordia (red arrows), and mis-patterned CNS neurons (compare red double arrows and brackets). MO, morpholino. **b**, Knockdown outcomes of seven genes with variants present in microcephaly patients: interocular measurements of bright-field images from control and loss-of-function embryos reveal significant decreases in head size. A neuronal antibody stain (anti-HuC/D, green channel) labels the brains of control and morphant zebrafish. Measurements taken across the widest extent of the midbrain identify significant reductions in brain size, probably underlying the concomitant head-size reductions seen in bright-field images. In **b**, tables show average percentage reduction in head and brain width, and *P* values of a *t*-test. Original magnifications: **a**, 5× (*pkn2a*), 10× (*fryl* and *psmd3*, bright-field) and 40× (psmd3, green channel); **b**, 10× (brightfield) and 20× (green channel).

phenotypes which could not be meaningfully linked to patient phenotypes. As many of our candidate developmental disorder genes carried heterozygous loss-of-function variants (*de novo* mutations), it is to be expected that the severity of loss-of-function phenotypes in zebrafish may exceed that observed in our patient cohort. The genes with proven non-redundant developmental roles can reasonably be assigned higher priority for downstream functional investigations and genetic analyses.

Our systematic review of gene perturbation in multiple species sought both confirmatory and contradictory (for example, healthy homozygous knockout) evidence from other animal models for these 21 apparently developmentally important genes. We identified 16 genes with solely confirmatory data, often from multiple different organisms, none with solely contradictory data, two with both confirmatory and contradictory evidence, and three with no evidence either way (Supplementary Table 8).

In summary, our analyses validate a large-scale, genotype-driven strategy for novel developmental-disorder-linked gene discovery that is complementary to the traditional phenotype-driven strategy of studying patients with very similar presentations, and is particularly effective for discovering novel developmental disorders with highly variable or indistinct clinical presentations. Our meta-analysis with previously published developmental disorder studies increased power to detect novel developmental-disorder-linked genes and highlights the shared genetic aetiologies between diverse neurodevelopmental disorders such as intellectual disability, epilepsy, autism and schizophrenia[22]. We identified significantly more pathogenic autosomal *de novo* mutations in females compared to males. An increased burden of monogenic disease among females with neurodevelopmental disorders has become more apparent[23,24], and our observations strengthen this proposition. Further investigations are required to assess whether males might be enriched for poly/oligogenic causation.

The 35 patients with pathogenic mutations in the 12 novel developmental-disorder-linked genes we discovered increased our diagnostic yield from 28% to 31%. This raises the question of what are the causes of the developmental disorders in the other 69% of patients. The undiagnosed patients are not obviously less severely affected than the diagnosed patients (for example, fewer phenotype terms, older age of recruitment). We anticipate that there are many more pathogenic, monogenic, coding mutations in these undiagnosed patients that we have detected, but for which compelling evidence is currently lacking. This hypothesis is supported by four strands of evidence: (1) modelling statistical power suggests that studying ~1,000 trios has only 5–10% power to detect an averagely mutable haploinsufficient developmental-disorder-linked gene (Extended Data Fig. 6a and Supplementary Information); (2) the expectation that our power to detect novel developmental-disorder-linked genes that operate recessively or by gain-of-function mechanisms will be lower than for haplosufficient genes; (3) the significant enrichment in undiagnosed patients of functional mutations in genes predicted to exhibit haploinsufficiency (Extended Data Fig. 6b); and (4) the strong enrichment for developmental phenotypes in the zebrafish knockdown screen.

Given our limited power to detect pathogenic mutations that act through dominant-negative or activating mechanisms, it was notable that in four of our novel genes (*COL4A3BP*, *PPP2R1A*, *PPP2R5D* and *PCGF2*) we observed identical *de novo* mutations in unrelated trios. Two hypotheses might explain this observation. First, that there is a vast number of different gain-of-function mutations, of which we are just scratching the surface in this study, or second, that these particular variants are enriched in our cohort owing to these mutations conferring a positive selective advantage in the germ line[25]. Analysis of larger data sets will be required to assess these hypotheses, although they are not necessarily mutually exclusive.

These considerations of the limited power of even nationwide studies such as ours motivate the international sharing of minimal genotypic and phenotypic data, for example through the DECIPHER web portal (http://decipher.sanger.ac.uk), to provide diagnoses for patients who would otherwise remain undiagnosed. Plausibly pathogenic variants observed in undiagnosed patients in our study (*de novo* SNVs, indels and CNVs, and biallelic loss of function in genes not yet associated with disease) are shared through DECIPHER, and we encourage other, comparable studies to adopt a similar approach.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 3 July; accepted 4 December 2014.
Published online 24 December 2014.

1. OMIM. *Online Mendelian Inheritance in Man* http://omim.org (2014).
2. Cooper, G. M. *et al.* A copy number variation morbidity map of developmental delay. *Nature Genet.* **43,** 838–846 (2011).

3. Allen, A. S. *et al. De novo* mutations in epileptic encephalopathies. *Nature* **501,** 217–221 (2013).
4. de Ligt, J. *et al.* Diagnostic exome sequencing in persons with severe intellectual disability. *N. Engl. J. Med.* **367,** 1921–1929 (2012).
5. Fromer, M. *et al. De novo* mutations in schizophrenia implicate synaptic networks. *Nature* **506,** 179–184 (2014).
6. Iossifov, I. *et al. De novo* gene disruptions in children on the autistic spectrum. *Neuron* **74,** 285–299 (2012).
7. Neale, B. M. *et al.* Patterns and rates of exonic *de novo* mutations in autism spectrum disorders. *Nature* **485,** 242–245 (2012).
8. O'Roak, B. J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations. *Nature* **485,** 246–250 (2012).
9. Rauch, A. *et al.* Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* **380,** 1674–1682 (2012).
10. Sanders, S. J. *et al. De novo* mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485,** 237–241 (2012).
11. Zaidi, S. *et al. De novo* mutations in histone-modifying genes in congenital heart disease. *Nature* **498,** 220–223 (2013).
12. King, D. A. *et al.* A novel method for detecting uniparental disomy from trio genotypes identifies a significant excess in children with developmental disorders. *Genome Res.* **24,** 673–687 (2014).
13. Wright, C. F. *et al.* Genetic diagnosis of developmental disorders in the DDD study: scalable analysis of genome-wide research data. *Lancet* http://dx.doi.org/10.1016/S0140-6736(14)61705-0 (2014).
14. Smith, B. H. *et al.* Cohort profile: Generation Scotland: Scottish Family Health Study (GS:SFHS). The study, its participants and their potential for genetic research on health and illness. *Int. J. Epidemiol.* **42,** 689–700 (2013).
15. Samocha, K. E. *et al.* A framework for the interpretation of *de novo* mutation in human disease. *Nature Genet.* **46,** 944–950 (2014).
16. Boumil, R. M. *et al.* A missense mutation in a highly conserved alternate exon of dynamin-1 causes epilepsy in fitful mice. *PLoS Genet.* **6,** e1001046 (2010).
17. Kumagai, K., Kawano, M., Shinkai-Ouchi, F., Nishijima, M. & Hanada, K. Interorganelle trafficking of ceramide is regulated by phosphorylation-dependent cooperativity between the PH and START domains of CERT. *J. Biol. Chem.* **282,** 17758–17766 (2007).
18. Walter, G. & Ruediger, R. Mouse model for probing tumor suppressor activity of protein phosphatase 2A in diverse signaling pathways. *Cell Cycle* **11,** 451–459 (2012).
19. Lim, E. T. *et al.* Rare complete knockouts in humans: population distribution and significant role in autism spectrum disorders. *Neuron* **77,** 235–242 (2013).
20. Kettleborough, R. N. *et al.* A systematic genome-wide analysis of zebrafish protein-coding gene function. *Nature* **496,** 494–497 (2013).
21. Pickart, M. A. *et al.* Genome-wide reverse genetics framework to identify novel functions of the vertebrate secretome. *PLoS ONE* **1,** e104 (2006).
22. Craddock, N. & Owen, M. J. The Kraepelinian dichotomy—going, going... but still not gone. *Br. J. Psychiatry* **196,** 92–95 (2010).
23. Jacquemont, S. *et al.* A higher mutational burden in females supports a ''female protective model'' in neurodevelopmental disorders. *Am. J. Hum. Genet.* **94,** 415–425 (2014).
24. Levy, D. *et al.* Rare *de novo* and transmitted copy-number variation in autistic spectrum disorders. *Neuron* **70,** 886–897 (2011).
25. Goriely, A. & Wilkie, A. O. Paternal age effect mutations and selfish spermatogonial selection: causes and consequences for human disease. *Am. J. Hum. Genet.* **90,** 175–200 (2012).
26. Huang, N., Lee, I., Marcotte, E. M. & Hurles, M. E. Characterising and predicting haploinsufficiency in the human genome. *PLoS Genet.* **6,** e1001154 (2010).
27. Helsmoortel, C. *et al.* A SWI/SNF-related autism syndrome caused by de novo mutations in ADNP. *Nature Genet.* **46,** 380–384 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements We dedicate this paper to J. Tolmie and L. Brueton for their unwavering and enthusiastic support of the Deciphering Developmental Disorders project. We thank the families for their participation and patience. We thank M. Daly and K. Samocha for access to unpublished mutation rate estimates. We are grateful to S. Saunders, D. Smedley, D. Conrad, A. Ramu and N. Huang for access to data and algorithms. We thank the UK National Blood Service and the Generation Scotland: Scottish Family Health Study for access to DNA from controls. Generation Scotland has received core funding from the Chief Scientist Office of the Scottish Government Health Directorates CZD/16/6 and the Scottish Funding Council HR03006. The Deciphering Developmental Disorders study presents independent research commissioned by the Health Innovation Challenge Fund (grant number HICF-1009-003), a parallel funding partnership between the Wellcome Trust and the Department of Health, and the Wellcome Trust Sanger Institute (grant number WT098051). The views expressed in this publication are those of the author(s) and not necessarily those of the Wellcome Trust or the Department of Health. The study has UK Research Ethics Committee approval (10/H0305/83, granted by the Cambridge South REC, and GEN/284/12 granted by the Republic of Ireland REC). The research team acknowledges the support of the National Institute for Health Research, through the Comprehensive Clinical Research Network.

Author Contributions See Supplementary Information for author contribution details.

Author Information Data can be accessed at the European Genome Phenome Archive (https://www.ebi.ac.uk/ega/) under accession number EGAS00001000775. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.E.H. (meh@sanger.ac.uk).

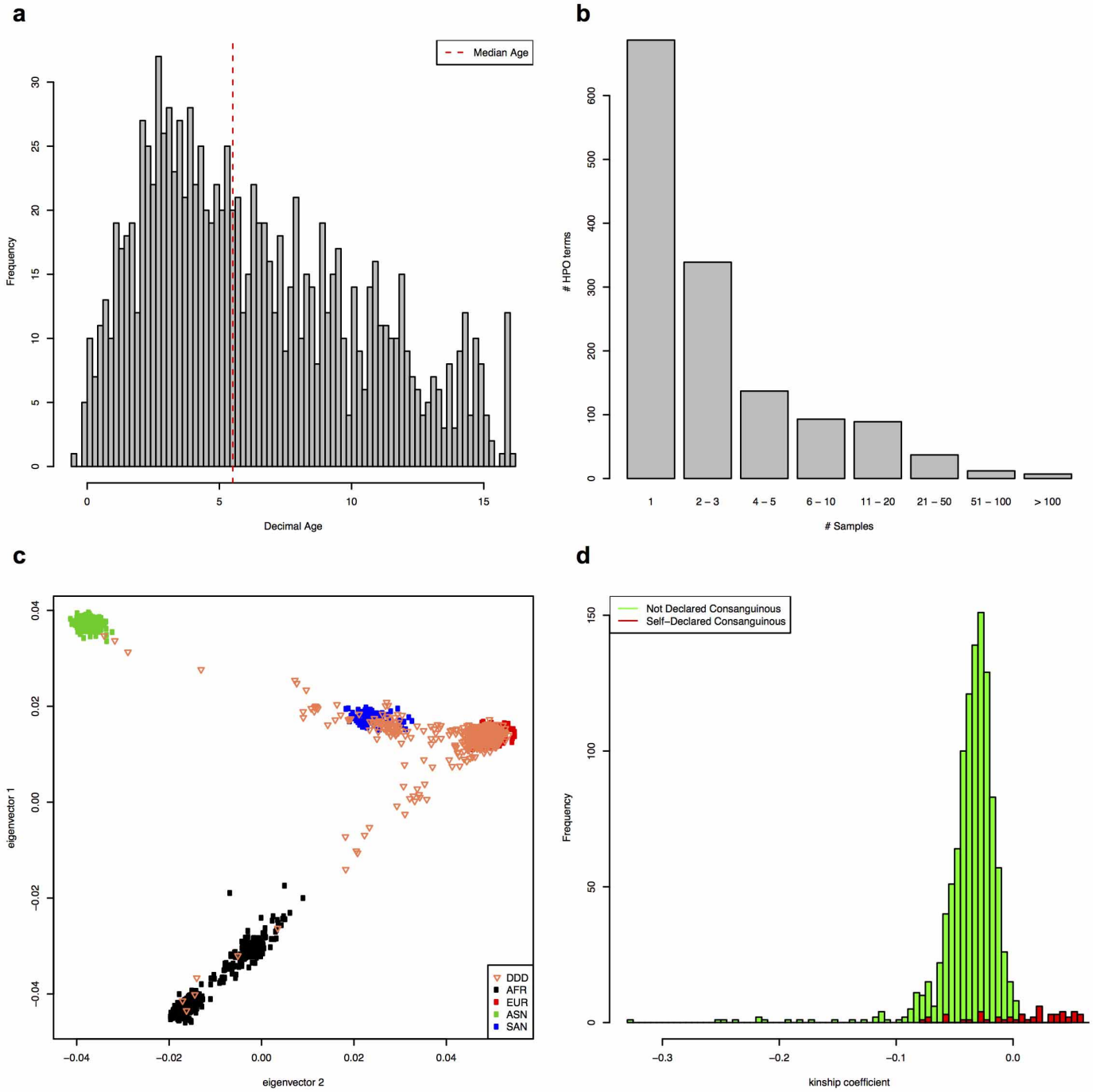
The Deciphering Developmental Disorders Study

T. W. Fitzgerald[1]*, S. S. Gerety[1]*, W. D. Jones[1]*, M. van Kogelenberg[1]*, D. A. King[1], J. McRae[1], K. I. Morley[1], V. Parthiban[1], S. Al-Turki[1], K. Ambridge[1], D. M. Barrett[1], T. Bayzetinova[1], S. Clayton[1], E. L. Coomber[1], S. Gribble[1], P. Jones[1], N. Krishnappa[1], L. E. Mason[1], A. Middleton[1], R. Miller[1], E. Prigmore[1], D. Rajan[1], A. Sifrim[1], A. R. Tivey[1], M. Ahmed[2,3,4], N. Akawi[1], R. Andrews[1], U. Anjum[5], H. Archer[6,7], R. Armstrong[8], M. Balasubramanian[9], R. Banerjee[1], D. Baralle[2,3,4], P. Batstone[10], D. Baty[11], C. Bennett[12], J. Berg[11], B. Bernhard[13], A. P. Bevan[1], E. Blair[14], M. Blyth[12], D. Bohanna[15], L. Bourdon[13], D. Bourn[16], A. Brady[13], E. Bragin[1], C. Brewer[17], L. Brueton[15], K. Brunstrom[18], S. J. Bumpstead[1], D. J. Bunyan[2,3,4], J. Burn[16], J. Burton[1], N. Canham[13], B. Castle[17], K. Chandler[19], S. Clasper[14], J. Clayton-Smith[19], T. Cole[15], A. Collins[2,3,4], M. N. Collinson[2,3,4], F. Connell[20], N. Cooper[15], H. Cox[15], L. Cresswell[21], G. Cross[22], Y. Crow[19], M. D'Alessandro[10], T. Dabir[23], R. Davidson[24], S. Davies[6,7], J. Dean[10], C. Deshpande[20], G. Devlin[17], A. Dixit[22], A. Dominiczak[25], C. Donnelly[19], D. Donnelly[23], A. Douglas[26], A. Duncan[24], J. Eason[22], S. Edkins[1], S. Ellard[17], P. Ellis[1], F. Elmslie[5], K. Evans[6,7], S. Everest[17], T. Fendick[20], R. Fisher[16], F. Flinter[20], N. Foulds[2,3,4], A. Fryer[26], B. Fu[1], C. Gardiner[24], L. Gaunt[19], N. Ghali[13], R. Gibbons[14], S. L. Gomes Pereira[1], J. Goodship[16], D. Goudie[11], E. Gray[1], P. Greene[27], L. Greenhalgh[26], L. Harrison[2,3,4], R. Hawkins[28], S. Hellens[16], A. Henderson[16], E. Hobson[12], S. Holden[8], S. Holder[13], G. Hollingsworth[18], T. Homfray[5], M. Humphreys[23], J. Hurst[18], S. Ingram[9], M. Irving[20], J. Jarvis[15], L. Jenkins[18], D. Johnson[9], D. Jones[1], E. Jones[19], D. Josifova[20], S. Joss[23], B. Kaemba[21], S. Kazembe[21], B. Kerr[19], U. Kini[14], E. Kinning[24], G. Kirby[15], C. Kirk[23], E. Kivuva[17], A. Kraus[12], D. Kumar[6,7], K. Lachlan[2,3,4], W. Lam[27], A. Lampe[27], C. Langman[20], M. Lees[18], D. Lim[15], G. Lowther[24], S. A. Lynch[29], A. Magee[23], E. Maher[27], S. Mansour[5], K. Marks[5], K. Martin[22], U. Maye[26], E. McCann[6,7], V. McConnell[23], M. McEntagart[5], R. McGowan[10], K. McKay[15], S. McKee[23], D. J. McMullan[15], S. McNerlan[23], S. Mehta[8], K. Metcalfe[19], E. Miles[19], S. Mohammed[20], T. Montgomery[16], D. Moore[27], S. Morgan[6,7], A. Morris[30], J. Morton[15], H. Mugalaasi[6,7], V. Murday[24], L. Nevitt[9], R. Newbury-Ecob[28], A. Norman[15], R. O'Shea[29], C. Ogilvie[20], S. Park[8], M. J. Parker[9], C. Patel[15], J. Paterson[8], S. Payne[13], J. Phipps[14], D. T. Pilz[6,7], D. Porteous[31], N. Pratt[11], K. Prescott[12], S. Price[14], A. Pridham[14], A. Procter[6,7], H. Purnell[14], N. Ragge[15], J. Rankin[17], L. Raymond[8], D. Rice[11], L. Robert[20], E. Roberts[28], G. Roberts[26], J. Roberts[8], P. Roberts[12], A. Ross[10], E. Rosser[18], A. Saggar[5], S. Samant[10], R. Sandford[8], A. Sarkar[22], S. Schweiger[11], C. Scott[1], R. Scott[18], A. Selby[22], A. Seller[14], C. Sequeira[13], N. Shannon[22], S. Sharif[15], C. Shaw-Smith[17], E. Shearing[9], D. Shears[14], I. Simonic[8], D. Simpkin[1], R. Singzon[13], Z. Skitt[19], A. Smith[12], B. Smith[32], K. Smith[9], S. Smithson[28], L. Sneddon[16], M. Splitt[16], M. Squires[12], F. Stewart[23], H. Stewart[14], M. Suri[22], V. Sutton[26], G. J. Swaminathan[1], E. Sweeney[26], K. Tatton-Brown[5], C. Taylor[9], R. Taylor[5], M. Tein[15], I. K. Temple[2,3,4], J. Thomson[12], J. Tolmie[24], A. Torokwa[2,3,4], B. Treacy[8], C. Turner[17], P. Turnpenny[17], C. Tysoe[17], A. Vandersteen[13], P. Vasudevan[21], J. Vogt[15], E. Wakeling[13], D. Walker[1], J. Waters[18], A. Weber[26], D. Wellesley[2,3,4], M. Whiteford[24], S. Widaa[1], S. Wilcox[8], D. Williams[15], N. Williams[24], G. Woods[8], C. Wragg[28], M. Wright[16], F. Yang[1], M. Yau[20], N. P. Carter[1], M. Parker[33], H. V. Firth[1,8], D. R. FitzPatrick[27], C. F. Wright[1], J. C. Barrett[1] & M. E. Hurles[1]

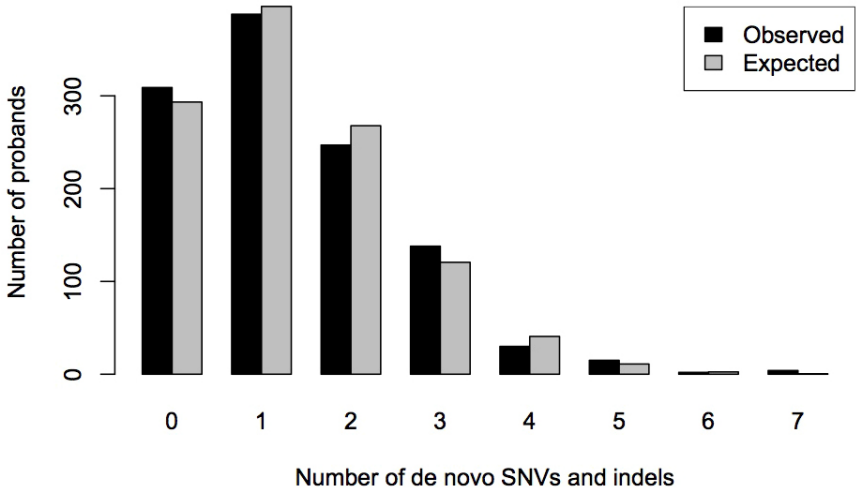
[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [2]Wessex Clinical Genetics Service, University Hospital Southampton, Princess Anne Hospital, Coxford Road, Southampton SO16 5YA, UK. [3]Wessex Regional Genetics Laboratory, Salisbury NHS Foundation Trust, Salisbury District Hospital, Odstock Road, Salisbury, Wiltshire SP2 8BJ, UK. [4]Faculty of Medicine, University of Southampton, Southampton SO16 6YD, UK. [5]South West Thames Regional Genetics Centre, St George's Healthcare NHS Trust, St George's, University of London, Cranmer Terrace, London SW17 0RE, UK. [6]Institute of Medical Genetics, University Hospital of Wales, Heath Park, Cardiff CF14 4XW, UK. [7]Department of Clinical Genetics, Block 12, Glan Clwyd Hospital, Rhyl, Denbighshire LL18 5UJ, UK. [8]East Anglian Medical Genetics Service, Box 134, Cambridge University Hospitals NHS Foundation Trust, Cambridge Biomedical Campus, Cambridge CB2 0QQ, UK. [9]Sheffield Regional Genetics Services, Sheffield Children's NHS Trust, Western Bank, Sheffield S10 2TH, UK. [10]North of Scotland Regional Genetics Service, NHS Grampian, Department of Medical Genetics Medical School, Foresterhill, Aberdeen AB25 2ZD, UK. [11]East of Scotland Regional Genetics Service, Human Genetics Unit, Pathology Department, NHS Tayside, Ninewells Hospital, Dundee DD1 9SY, UK. [12]Yorkshire Regional Genetics Service, Leeds Teaching Hospitals NHS Trust, Department of Clinical Genetics, Chapel Allerton Hospital, Chapeltown Road, Leeds LS7 4SA, UK. [13]North West Thames Regional Genetics Centre, North West London Hospitals NHS Trust, The Kennedy Galton Centre, Northwick Park And St Mark's NHS Trust Watford Road, Harrow HA1 3UJ, UK. [14]Oxford Regional Genetics Service, Oxford Radcliffe Hospitals NHS Trust, The Churchill Old Road, Oxford OX3 7LJ, UK. [15]West Midlands Regional Genetics Service, Birmingham Women's NHS Foundation Trust, Birmingham Women's Hospital, Edgbaston, Birmingham B15 2TG, UK. [16]Northern Genetics Service, Newcastle upon Tyne Hospitals NHS Foundation Trust, Institute of Human Genetics, International Centre for Life, Central Parkway, Newcastle upon Tyne NE1 3BZ, UK. [17]Peninsula Clinical Genetics Service, Royal Devon and Exeter NHS Foundation Trust, Clinical Genetics Department, Royal Devon & Exeter Hospital (Heavitree), Gladstone Road, Exeter EX1 2ED, UK. [18]North East Thames Regional Genetics Service, Great Ormond Street Hospital for Children NHS Foundation Trust, Great Ormond Street Hospital, Great Ormond Street, London WC1N 3JH, UK. [19]Manchester Centre for Genomic Medicine, St Mary's Hospital, Central Manchester University Hospitals NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester M13 9WL,

UK. [20]South East Thames Regional Genetics Centre, Guy's and St Thomas' NHS Foundation Trust, Guy's Hospital, Great Maze Pond, London SE1 9RT, UK. [21]Leicestershire Genetics Centre, University Hospitals of Leicester NHS Trust, Leicester Royal Infirmary (NHS Trust), Leicester LE1 5WW, UK. [22]Nottingham Regional Genetics Service, City Hospital Campus, Nottingham University Hospitals NHS Trust, The Gables, Hucknall Road, Nottingham NG5 1PB, UK. [23]Northern Ireland Regional Genetics Centre, Belfast Health and Social Care Trust, Belfast City Hospital, Lisburn Road, Belfast BT9 7AB, UK. [24]West of Scotland Regional Genetics Service, NHS Greater Glasgow and Clyde, Institute Of Medical Genetics, Yorkhill Hospital, Glasgow G3 8SJ, UK. [25]College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow G12 8QQ, UK. [26]Merseyside and Cheshire Genetics Service, Liverpool Women's NHS Foundation Trust, Department of Clinical Genetics, Royal Liverpool Children's Hospital Alder Hey, Eaton Road, Liverpool L12 2AP, UK. [27]MRC Human Genetics Unit, MRC IGMM, University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, UK. [28]Bristol Genetics Service (Avon, Somerset, Gloucs and West Wilts), University Hospitals Bristol NHS Foundation Trust, St Michael's Hospital, St Michael's Hill, Bristol BS2 8DT, UK. [29]National Centre for Medical Genetics, Our Lady's Children's Hospital, Crumlin, Dublin 12, Ireland. [30]School of Molecular, Genetic and Population Health Sciences, University of Edinburgh Medical School, Teviot Place, Edinburgh EH8 9AG, UK. [31]University of Edinburgh, Institute of Genetics & Molecular Medicine, Western General Hospital, Crewe Road South, Edinburgh EH4 2XU, UK. [32]School of Medicine, Dundee University, Mackenzie Building, Kirsty Semple Way, Ninewells Hospital and Medical School, Dundee DD2 4RB, UK. [33]The Ethox Centre, Nuffield Department of Population Health, University of Oxford, Old Road Campus, Oxford OX3 7LF, UK.

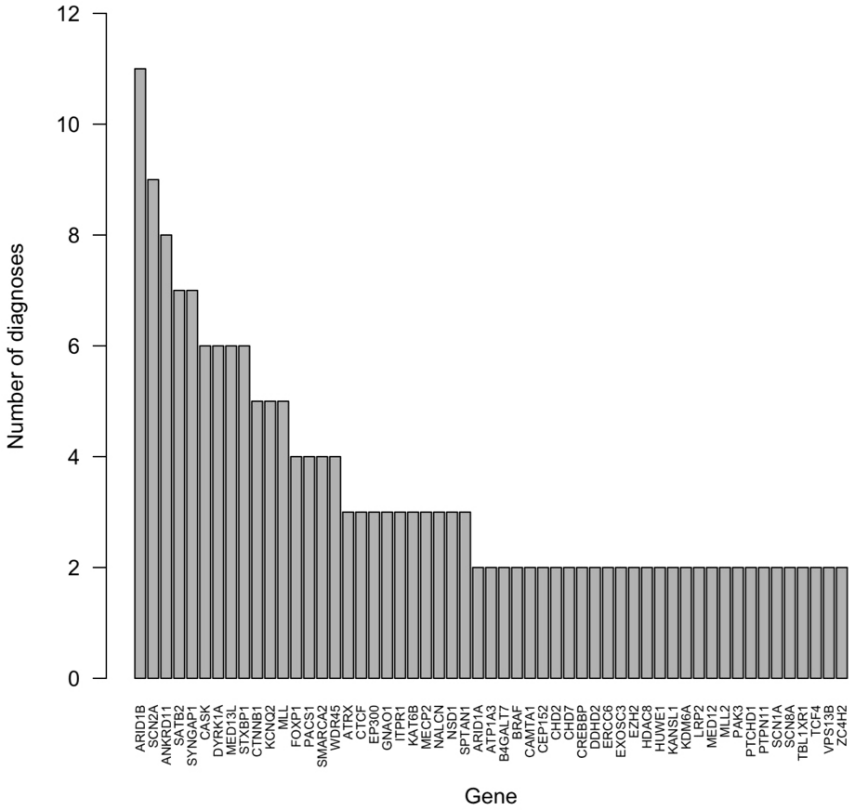
*These authors contributed equally to this work.



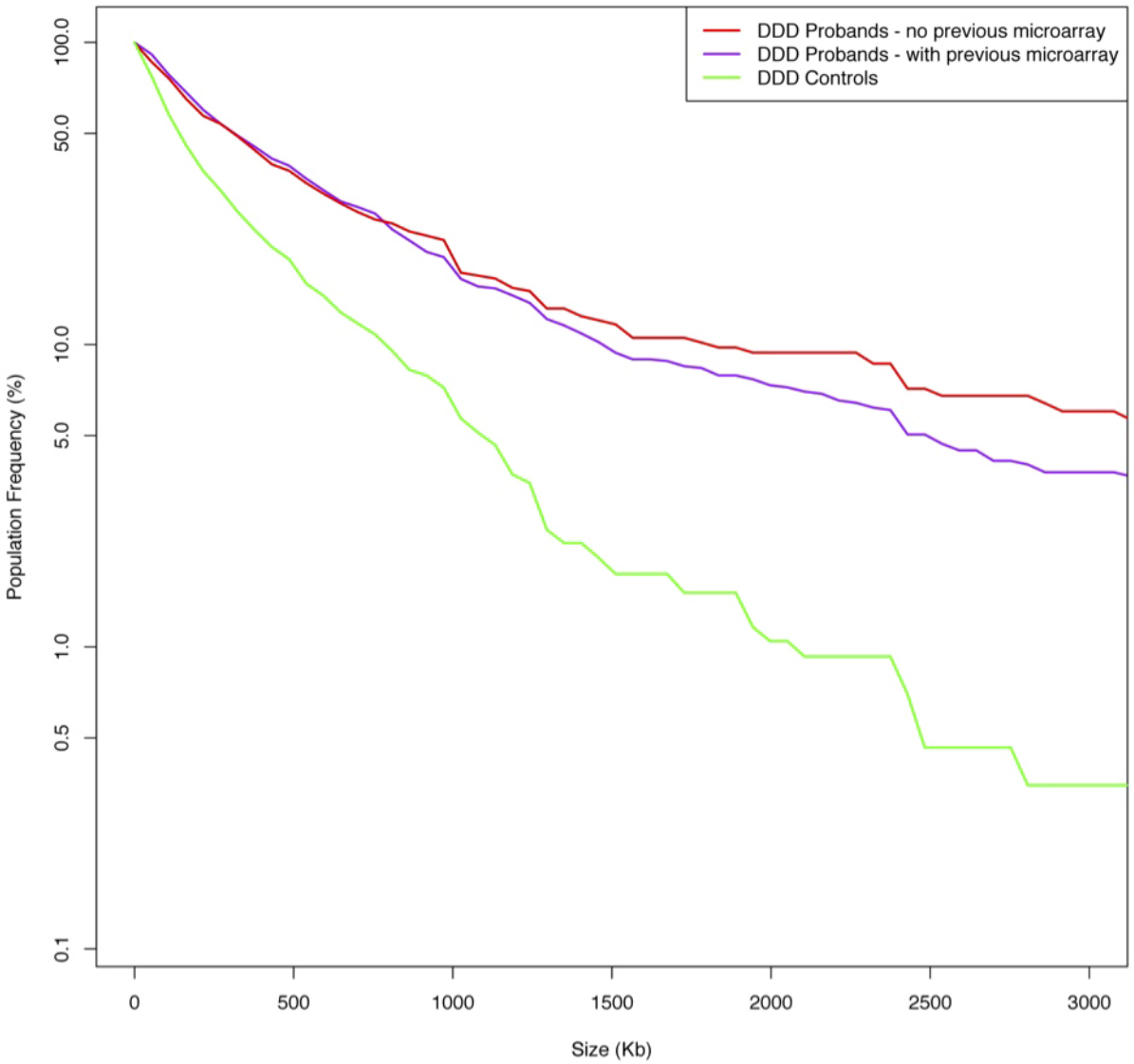
Extended Data Figure 1 | Characteristics of the families. **a**, Gestation-adjusted decimal age (years) at last clinical assessment. The histogram shows the distribution of the gestation-adjusted decimal age at last clinical assessment across the 1,133 probands. The dashed red line shows the median age. **b**, Frequency of human phenotype ontology (HPO) term usage. Bar plot showing, for each used HPO term, the number of times it was observed across the 1,133 proband patient records. **c**, Projection PCA plot of the 1,133 probands. PCA plot of 1,133 DDD probands projected onto a PCA analysis using four different HapMap populations from the 1000 genomes project. Black, African; red, European; green, east Asian; blue, south Asian; and the 1,133 DDD probands are represented by orange triangles. **d**, Self-declared and genetically defined consanguinity. Overlaid histogram showing the distribution of kinship coefficients from KING comparing parental samples for each trio. Green, trios where consanguinity was not entered in the patient record on DECIPHER; red, trios consanguinity was declared in the patient record on DECIPHER.



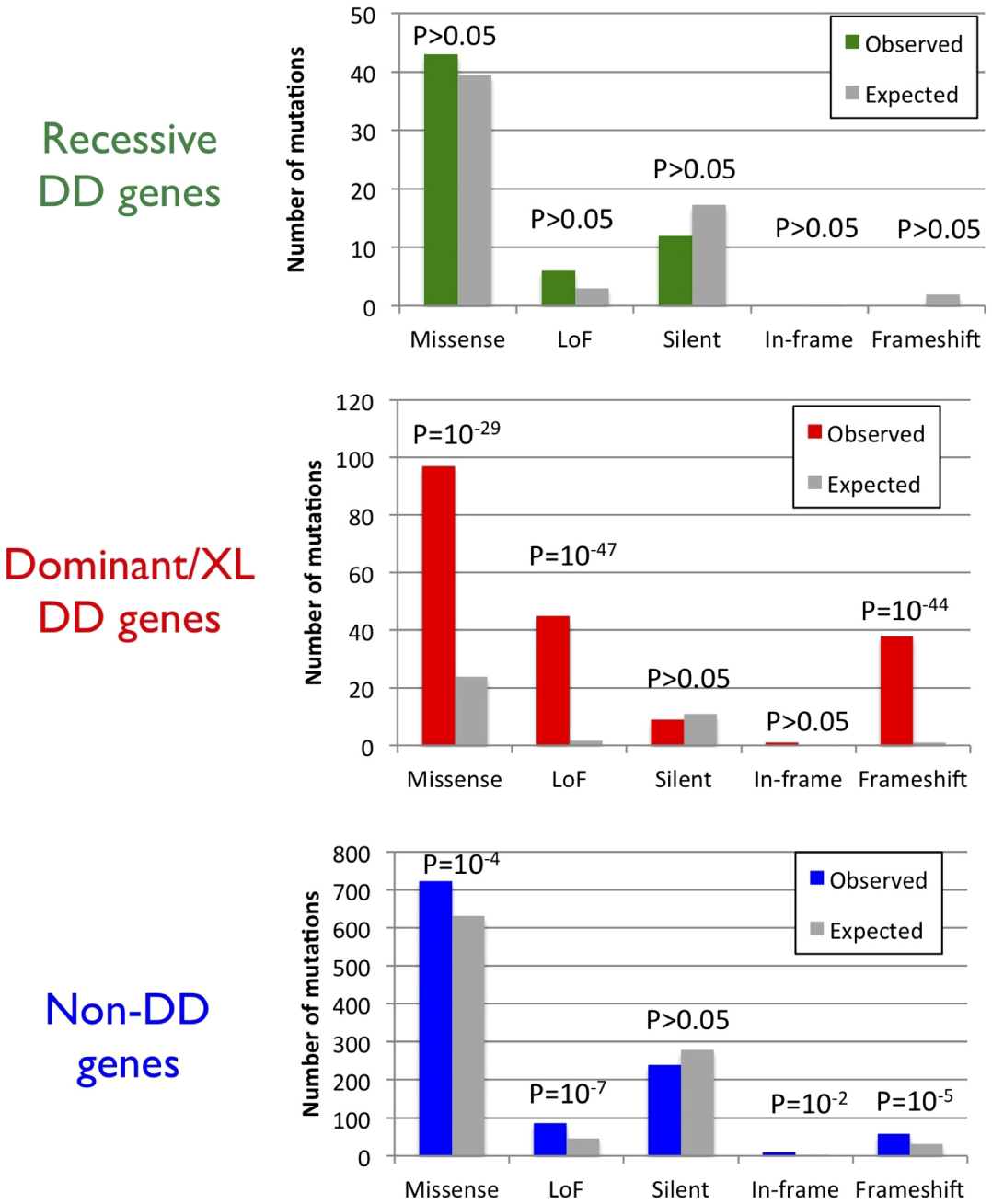
Extended Data Figure 2 | Number of validated *de novo* SNVs and indels per proband. Bar plot showing the distribution of the observed number of validated SNVs and indels per proband sample, and the expected distribution assuming a Poisson distribution with the same mean as the observed distribution.



Extended Data Figure 3 | Number of diagnoses per gene. Histogram showing the number of diagnoses per gene for genes with at least two diagnoses from different proband samples.

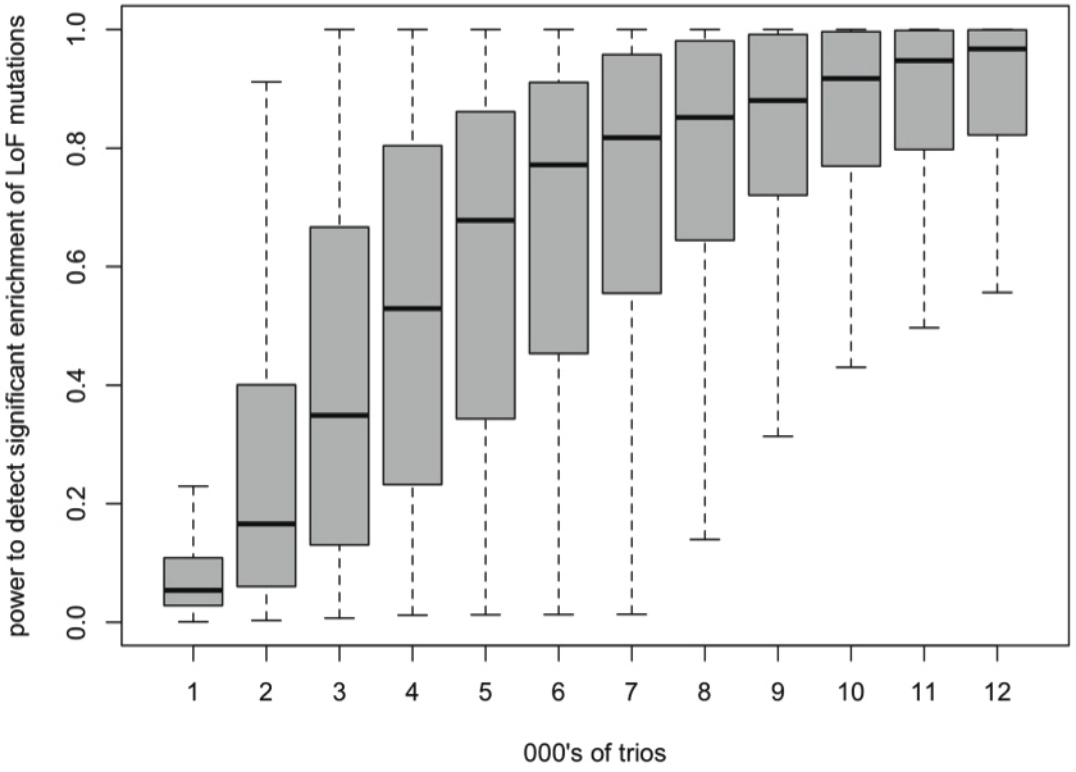


Extended Data Figure 4 | Burden of large CNVs in 1,133 DDD proband samples. Plot comparing the frequency of rare CNVs in three sample groups against CNV size. The *y* axis is on a log scale. Red, DDD probands who have not had previous microarray based genetic testing; purple, DDD probands who have had negative previous microarray-based genetic testing; green, DDD controls.

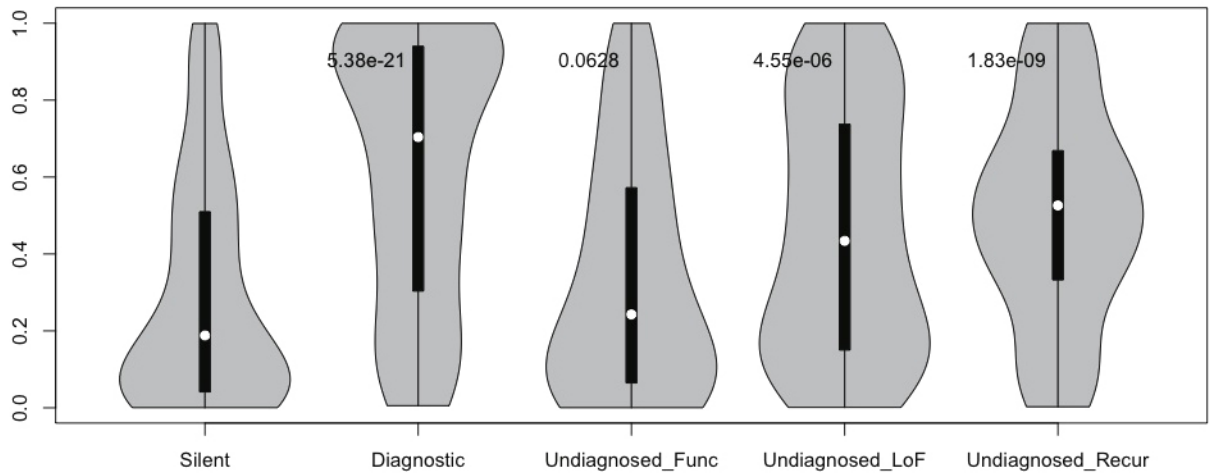


Extended Data Figure 5 | Expected and observed numbers of *de novo* mutations. The expected and observed numbers of mutations of different functional consequences in three mutually exclusive sets of genes are shown, along with the *P* value from an assessment of a statistical excess of observed mutations. The three classes of genes are described in the main text.

A



B



Extended Data Figure 6 | Haploinsufficiency analyses. a, Saturation analysis for detecting haploinsufficient developmental-disorder-linked genes. A box plot showing the distribution of statistical power to detect a significant enrichment of loss-of-function mutations across 18,272 genes in the genome, for different numbers of trios studied, from 1,000 trios to 12,000 trios. Line within the box shows the median, box shows the interquartile range and the whiskers show the most extreme values within 1.5 times the interquartile range from the box. **b**, Distribution of haploinsufficiency scores in selected sets of *de novo* mutations. Violin plot of haploinsufficiency scores in five sets of *de novo* mutations. Silent, all synonymous mutations; diagnostic, mutations in known developmental-disorder-linked genes in diagnosed individuals; undiagnosed_Func, all functional mutations in undiagnosed individuals; undiagnosed_LoF, all loss-of-function mutations in undiagnosed individuals; undiagnosed_recur, mutations in genes with recurrent functional mutations in undiagnosed individuals. *P* values for a Mann–Whitney *U*-test comparing each of the latter four distributions to that observed for the silent (synonymous) variants are plotted at the top of each violin. Dot indicates the median, box is interquartile range and whiskers are the most extreme values within 1.5 times the interquartile range from the box.

Extended Data Table 1 | Novel genes with suggestive evidence for a role in developmental disorder

Evidence	Gene	*de novos* DDD (Missense, LoF)	*de novos* Meta (Missense, LoF)	P Value	Test	Mutation Clustering	Predicted Haploinsufficiency
De novo enrichment + additional evidence	*NAA15*	1 (0,1)	3 (0,3)	1.64E-06	Meta	No	7.5%
	ZBTB20	3 (1,2)	3 (1,2)	4.84E-06	DDD	No	0.2%
	NAA10	2 (2,0)	3 (3,0)	8.28E-06	Meta	No	34.1%
	TRIP12	3 (1,2)	4 (2,2)	2.13E-05	Meta	No	3.8%
	USP9X	3 (1,2)	3 (1,2)	5.14E-05	DDD	No	3.8%
	KAT6A	2 (0,2)	2 (0,2)	7.91E-05	DDD	No	19.0%

Six genes with suggestive evidence to be novel developmental-disorder-linked genes. The number of unrelated patients with independent functional or loss-of-function mutations in the DDD cohort or the wider meta-analysis data set including DDD patients is listed. The *P* value reported is the minimum *P* value from the testing of the DDD data set and the meta-analysis data set. The data set that gave this minimal *P* value is also reported. Mutations are considered to be clustered if the *P* value of clustering of functional SNVs is less than 0.01. Predicted haploinsufficiency is reported as a percentile of all genes in the genome, with ~0% being highly likely to be haploinsufficient and 100% very unlikely to be haploinsufficient, based on the prediction score described in ref. 26 updated to enable predictions for a higher fraction of genes in the genome. *NAA10* is already known to cause an X-linked recessive developmental disorder in males, but here we identified missense mutations in females, suggesting a different, X-linked dominant, disorder.

Extended Data Table 2 | Biallelic loss of function and damaging functional variants

Biallelic Variant Types	Untransmitted Diplotypes (n=1080)	Likely Dominant Probands (n=270)	Other Probands (n=810)
LoF/LoF (Genome-wide)	110	17	86
LoF/Dam (Genome-wide)	87	21	71
Dam/Dam (Genome-wide)	312	90	264
LoF/LoF (DDG2P Biallelic)	1	1	3
LoF/Dam (DDG2P Biallelic)	2	0	6
Dam/Dam (DDG2P Biallelic)	26	7	25

Rare (MAF <5%) biallelic loss of function and damaging functional variants in uninherited diplotypes and probands. 'Likely dominant probands' refers to probands with a reported *de novo* mutation or affected parents, and 'other probands' refers to all remaining probands. 'DDG2P biallelic' refers to confirmed and probable DDG2P genes with a biallelic mode of inheritance. See Supplementary Methods for details of variant processing.

Extended Data Table 3 | Zebrafish modelling identifies 21 developmentally important candidate genes

Gene	# patients	Variant	Patient phenotypes	Phenotypic concordance	Relevant knockdown phenotypes
BTBD9	2/1	Biallelic LoF/*De novo* Missense	Seizures, microcephaly, hypertonia	Strong	Reduced head size, brain volume
CHD3	1/2	*De novo* LoF/Missense	CNS and craniofacial defects	Strong	Abnormal head shape
DDX3X	1/3	*De novo* LoF/Missense	Moderately short stature, microcephaly, CNS defects	Strong	Reduced head size, brain volume
ETF1	1	*De novo* LoF	CNS and craniofacial defects, seizures, microcephaly, hypertelorism	Strong	Reduced head size, brain volume
FRYL	1	*De novo* LoF	Short stature, craniofacial and cardiac defects	Strong	Cardiac defects, reduced axis length
PKN2	1	*De novo* Missense	CNS, cardiac, ear, and craniofacial defects, growth retardation	Strong	Cardiac, craniofacial cartilage, and growth defects
PSMD3	1	*De novo* Missense	Microcephaly, muscular hypotonia, seizures, growth abnormality	Strong	Reduced head size and neural defects
SCGN	1	Biallelic LoF	Seizures, microcephaly, CNS defects	Strong	Reduced head size, brain volume
SETD5	1	*De novo* LoF	Seizures, CNS and cardiac defects, poor motor coordination	Strong	Reduced head size, cardiac defects, abnormal locomotion
THNSL2	2	Biallelic LoF	Microcephaly, CNS and ear defects	Strong	Reduced head size, brain volume, neural defects
ZRANB1	2	*De novo* Missense	Microcephaly, muscle defects, seizures	Strong	Reduced head size and neural defects
DPEP2	1	Biallelic LoF	CNS defects, growth retardation	Moderate	Growth reduction
PSD2	1	*De novo* LoF	CNS defects, hypertonia, seizures	Moderate	Abnormal musculature, CNS and locomotion
SAP130	1	*De novo* LoF	Short stature, hypotonia, hypotelorism	Moderate	Abnormal locomotion
CNOT1	1/1	*De novo* LoF/Missense	Short stature, cardiac, CNS, ear and craniofacial defects	Weak	Multisystem
DTWD2	1	*De novo* LoF	CNS defects, seizures	Weak	Multisystem
ILVBL	1	*De novo* LoF	CNS and craniofacial defects	Weak	Multisystem
NONO	1	*De novo* LoF	CNS and ear defects, hypotonia, growth retardation	Weak	Multisystem, with otic and growth defects
POGZ	2	*De novo* LoF	CNS and ear defects, hypotonia, seizures, coloboma	Weak	Multisystem
SMARCD1	1/1	*De novo* LoF/Missense	CNS defects, hypotonia	Weak	Multisystem
WWC1	1	*De novo* Missense	CNS defects, hypertelorism	None	None

This table summarizes the 21 genes for which knockdown results in developmental phenotypes in zebrafish. The '# patients' column indicates how many patients were identified as carrying variants in these genes. Split numbers indicate the breakdown of variant types (for example, for *BTBD9*, 2/1 is two biallelic loss of function and one *de novo* missense carrying patients). A summary of the patient phenotypes is listed, as well as the relevant phenotypes observed in zebrafish knockdown experiments. Phenotypic concordance categories indicate the degree of overlap between the zebrafish phenotyping and the patient phenotypes. Weak concordance typically is the result of severe, multisystem phenotypes in zebrafish. See Supplementary Information for more detailed phenotype information.

LETTER

doi:10.1038/nature14103

Orientation columns in the mouse superior colliculus

Evan H. Feinberg[1] & Markus Meister[1,2]

More than twenty types of retinal ganglion cells conduct visual information from the eye to the rest of the brain[1,2]. Each retinal ganglion cell type tessellates the retina in a regular mosaic, so that every point in visual space is processed for visual primitives such as contrast and motion[3]. This information flows to two principal brain centres: the visual cortex and the superior colliculus. The superior colliculus plays an evolutionarily conserved role in visual behaviours[4], but its functional architecture is poorly understood. Here we report on population recordings of visual responses from neurons in the mouse superior colliculus. Many neurons respond preferentially to lines of a certain orientation or movement axis. We show that cells with similar orientation preferences form large patches that span the vertical thickness of the retinorecipient layers. This organization is strikingly different from the randomly interspersed orientation preferences in the mouse's visual cortex[5]; instead, it resembles the orientation columns observed in the visual cortices of large mammals[6–8]. Notably, adjacent superior colliculus orientation columns have only limited receptive field overlap. This is in contrast to the organization of visual cortex, where each point in the visual field activates neurons with all preferred orientations[9]. Instead, the superior colliculus favours specific contour orientations within ~30° regions of the visual field, a finding with implications for behavioural responses mediated by this brain centre.

We exposed the mouse superior colliculus (SC) for chronic brain imaging while leaving cortex intact (see Methods; Fig. 1a–c and Extended Data Fig. 1) and delivered the calcium indicator GCaMP6s as a neural activity reporter[10]. Awake mice head-fixed on a circular treadmill viewed stimuli on a tangent screen, while neuronal responses were monitored by two-photon microscopy (Fig. 1d). The focal plane of the microscope was roughly parallel to the surface of the SC and its retinotopic map of visual space[11,12]. The screen displayed thin bars drifting along their short axes (Fig. 1d), a stimulus that elicits orientation- or axis-tuned responses from many cells in the SC of anaesthetized mice, albeit by mechanisms that appear distinct from those of cortical neurons[13,14]. The animals remained stationary on most stimulus trials (72%, $n = 5$ animals, 208 stimulus blocks), and the fraction of stationary trials did not vary across visual stimuli ($P = 0.99$, Kruskal–Wallis test). Consequently, we report measurements from all trials regardless of locomotion; these results differ only subtly from those obtained by excluding running trials.

Neurons in the upper layers of the SC responded to drifting bars with large and reproducible transients in fluorescence that were often stronger to certain bar orientations than to others, consistent with previous reports of orientation tuning[14] (Fig. 1e, f). Unexpectedly, neighbouring neurons frequently displayed remarkably similar response profiles (Fig. 2a, b). Volumes of the SC (150 μm (anterior–posterior) × 280 μm (medial–lateral) × 40–80 μm (dorsal–ventral)) were analysed in several animals, with a focus on fields of view containing multiple groups of cells with different preferred orientations (Fig. 2c). The preferred orientations of orientation-selective cells separated horizontally by short distances (<100 μm) were much more alike than expected by chance (Fig. 2d and Extended Data Fig. 4a–c; mean $\Delta\theta \pm$ s.e.m., 28.6° ± 0.7; median $\Delta\theta$, 20.5° for 1,139 cell pairs), whereas preferred orientations of cells separated by greater distances (150–250 μm) were much less alike than expected by chance (Fig. 2d and Extended Data Fig. 4a–c; mean $\Delta\theta \pm$ s.e.m., 57.8° ± 1.2; median $\Delta\theta$, 63.6° for 440 cell pairs).

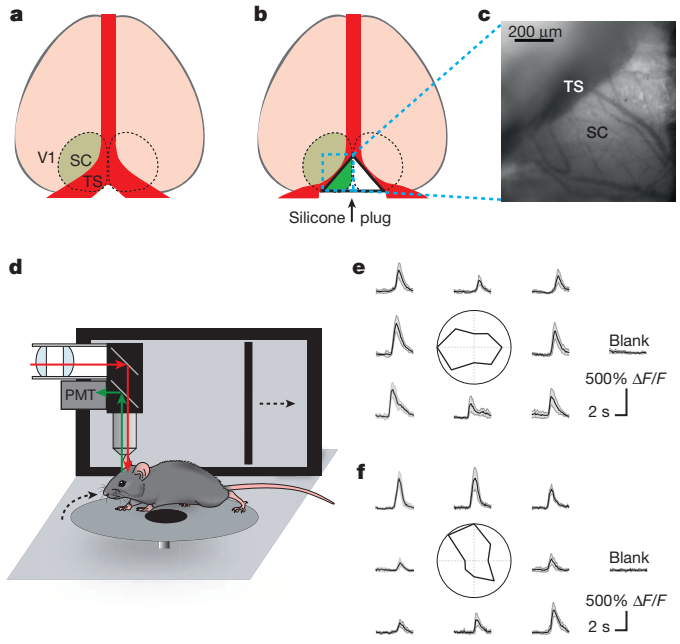


Figure 1 | Calcium imaging in awake mouse superior colliculus reveals orientation tuning. **a**, Schematic of mouse cerebral anatomy. The SC lies beneath visual cortex (V1) and the transverse sinus (TS). GCaMP6s-expressing SC is labelled in green. **b**, Schematic of cerebral anatomy after insertion of a triangular plug to reveal ~15–25% of the SC. **c**, Exposed portion of the SC. **d**, Schematic of experimental setup. Mice are head-fixed on a turntable and free to run. Visual stimuli are presented on a tangent screen while two-photon calcium imaging is used to record neural population activity. PMT, photomultiplier tube. **e**, **f**, Average fluorescence signal $\Delta F/F \pm$ s.d. of two SC neurons to 7 repetitions each of 8 directions of bar motion or a blank screen. Insets are polar plots of the peak responses.

[1]Center for Brain Science, Department of Molecular and Cellular Biology, Harvard University, 52 Oxford Street, Cambridge, Massachusetts 02138, USA. [2]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, California 91125, USA.

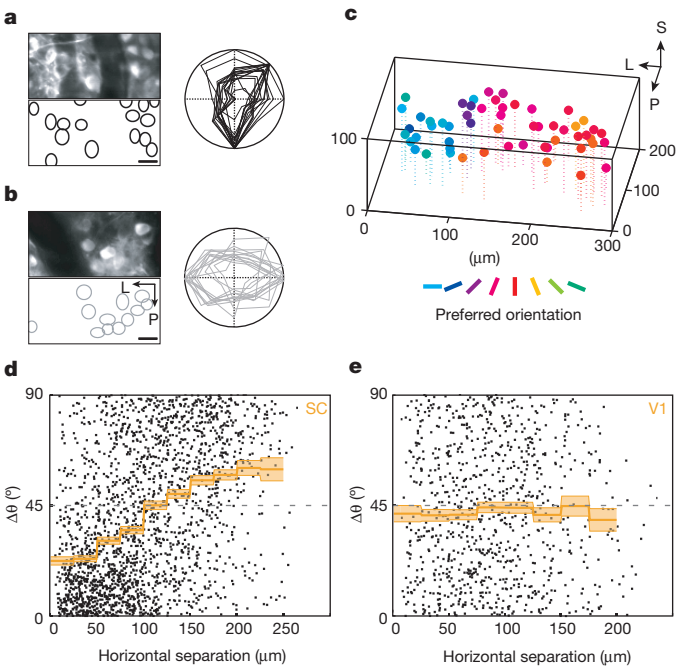


Figure 2 | Patches of neurons with similar orientation tuning in the superior colliculus. **a**, **b**, Two fields of cells in the SC (upper left panels) and corresponding elliptical regions of interest (lower left panels). Responses of the cells to drifting bars are overlaid in polar plots (right panels), normalized to peak responses. Scale bar, 20 μm. **c**, All orientation-tuned cells within a volume in the SC are plotted as spheres and colour-coded according to preferred orientations. L, lateral; P, posterior; S, superficial. **d**, **e**, Absolute value of the difference in preferred orientations plotted against horizontal distance in the SC (**d**) for 7 volumes in 6 animals ($n = 269$ cells, 2,077 pairs) and in V1 (**e**) for 3 volumes in 3 animals ($n = 104$ cells, 890 pairs). 0 and 90° correspond to identical and orthogonal orientation preferences, respectively, whereas 45° would be expected by chance (dashed grey line). Orange lines, means for 25-μm bins ± s.e.m.

These results were unexpected because the input to the SC from the retina is not thought to carry an orientation bias, even though individual retinal ganglion cells can be orientation-tuned[15]. We first considered the effects of optical projection from the flat tangent screen, which can alter the apparent width of a bar depending on its orientation, but the preferred orientations were not consistently biased towards or away from radial orientations (Extended Data Figs 1 and 2 and Supplementary Discussion). To confirm that the effects observed were not artefacts of our experimental system, we modified the surgical procedure to deliver GCaMP6s to both the SC and primary visual cortex (V1) and visualize both areas. Drifting bar stimuli elicited orientation-tuned responses from neurons in V1 (Extended Data Fig. 3) that were often sharper than in the SC[14,16] (mean orientation selectivity index (see Methods) of orientation-tuned neurons: 0.39 (V1) vs 0.31 (SC), $P = 0.003$), suggesting that surgical exposure of the SC spares V1 function. However, unlike in the SC, the arrangement of V1 neurons bore no relationship to their preferred orientation and was indistinguishable from chance[5]. (Fig. 2e and Extended Data Fig. 4d–f). Moreover, SC and V1 neurons had overlapping receptive fields, suggesting that the pattern observed in the SC is likely not inherited from the retina. This side-by-side comparison indicates that the orientation patches in the SC are not artefacts of the stimulus or imaging paradigms. Instead, the organization of orientation selectivity in the mouse SC differs substantially from that in visual cortex.

To probe the functional organization of the SC perpendicular to the brain surface we imaged neural responses at different depths. Pairs of neurons separated by <25 μm horizontally tended to share preferred orientations, regardless of depth separation, at least over 80 μm depth (Fig. 3a, b). In the deeper SC, fluorescence signals often became dimmer and less sharp, precluding efficient motion correction and accurate correction for neuropil contamination for single cells (Methods). Because SC cells and the surrounding neuropil are tuned alike (Extended Data Fig. 5), bulk fluorescence signals offer useful proxies for local orientation tuning that allowed analysis of deeper volumes (Fig. 3c). Orientation tuning was similar between slices of the same vertical column over depth separations up to 260 μm, but significantly different for slices drawn from different columns (Fig. 3d; $P < 0.001$, Kruskal–Wallis test). These results suggest that vertical columns of cells with similar orientation preferences span the retinorecipient SC (Fig. 3c, d).

To generate larger maps of orientation tuning in the SC, we turned to a complementary wide-field method: optical imaging of intrinsic signals[8]. Because most cells preferred orientations close to the cardinal axes (Extended Data Fig. 6), horizontal and vertical bars were used for intrinsic imaging experiments. Large patches (>200 μm diameter) of the SC preferred either horizontal or vertical bars (Fig. 4a). This arrangement was grossly reproducible across animals. The medial and lateral parts of the exposed SC, corresponding to the superior visual field and elevations close to the horizon, respectively, tended to prefer vertical bars—whereas the intervening area, which surveys intermediate elevations, tended to prefer horizontal bars (Fig. 4a–c). The orientation maps showed no relationship to the distortions introduced by the flat screen, and were robust to variations in the widths and velocities of the bars (Extended Data Fig. 7). These patches were reminiscent of the patterns observed in visual cortices of other mammalian species[7,8]. Attempts were made with limited success to aspirate the overlying cortex and blood vessels to expose more of the SC. In one instance, a small patch extending further anterior and lateral could be imaged, and the stereotyped orientation patches appeared in the expected locations.

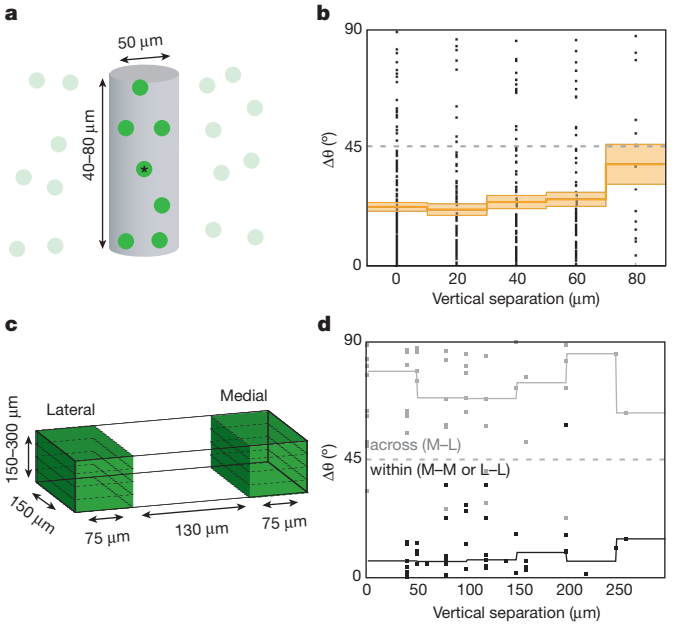


Figure 3 | Orientation patches form vertical columns. **a**, A cylinder centred on each neuron was projected through the volume and the similarity of its preferred orientation to that of each cell in the cylinder was determined. **b**, Difference in preferred orientations plotted against vertical distance. Orange lines, means for 20-μm bins ± s.e.m. Dashed grey line indicates chance. Differences across depths were not significant ($P > 0.1$, Kruskal–Wallis test; $n = 5$ animals, 397 pairs). **c**, Signals are averaged within 75 × 150 μm slices on the medial and lateral edges of an image plane, separated horizontally by 130 μm, at several depths along the vertical axis. **d**, Difference in preferred orientations plotted against vertical distance for orientation-tuned slices along the same vertical column (black dots) or adjacent vertical columns (grey dots). Lines indicate medians for 50-μm bins. M, medial slice; L, lateral slice. Data from 4 mice.

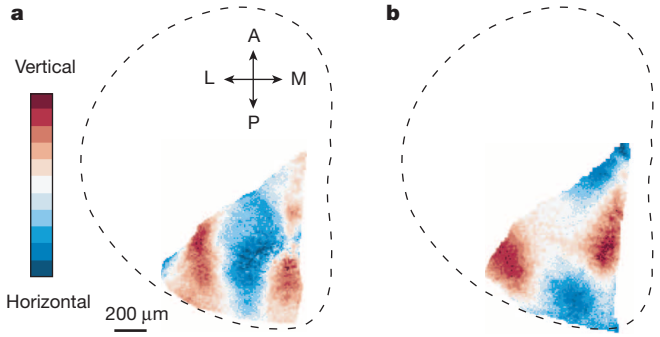


Figure 4 | Intrinsic imaging reveals larger orientation maps. **a**, Orientation map of the SC. Red areas respond more strongly to vertical bars and blue areas respond more strongly to horizontal bars. A, anterior; L, lateral; M, medial; P, posterior. The reflectance change $\Delta R/R$ from red to blue is 6×10^{-4} in **a** and **c**, 4×10^{-4} in **b** and 3×10^{-4} in **d**. The dashed outline indicates rough dimensions of the SC and approximate location of field of view. Midline blood vessels obscure the medial SC, containing cells with the most superior receptive fields. **b**, **c**, Orientation maps for two additional animals. **d**, Orientation map in an animal in which visual cortex and the transverse sinus had been surgically ablated weeks earlier. Granulation tissue over much of the SC limited the field of view.

This suggests that orientation columns in the SC arise even without input from the visual cortex (Fig. 4d).

These results indicate that cells in the SC with similar orientation preferences form patches within the retinotopic map of visual space, an arrangement that might preclude uniform coverage of different contour orientations throughout the visual field. This issue has been addressed in species with orientation columns in the visual cortex. There a uniform coverage is preserved because the grain of the orientation patches is so fine that any point in the visual field activates neurons of all possible orientation preferences[17–19]. To examine whether this applies in the mouse SC, we measured the projective fields on the SC for localized stimuli (Fig. 5a). Thin gratings produced activation stripes consistent with the known retinotopic map (Fig. 5b–d). Stripes had an average width of 170–190 μm (full-width at half-maximum, 4 stripes, 2 animals). Gratings separated by ~12° in space excited largely non-overlapping areas on the SC, with peak-to-peak distances of 100–170 μm ($n = 4$ pairs, 2 mice; Fig. 5b–d and Extended Data Figs 7 and 8). By comparison, the

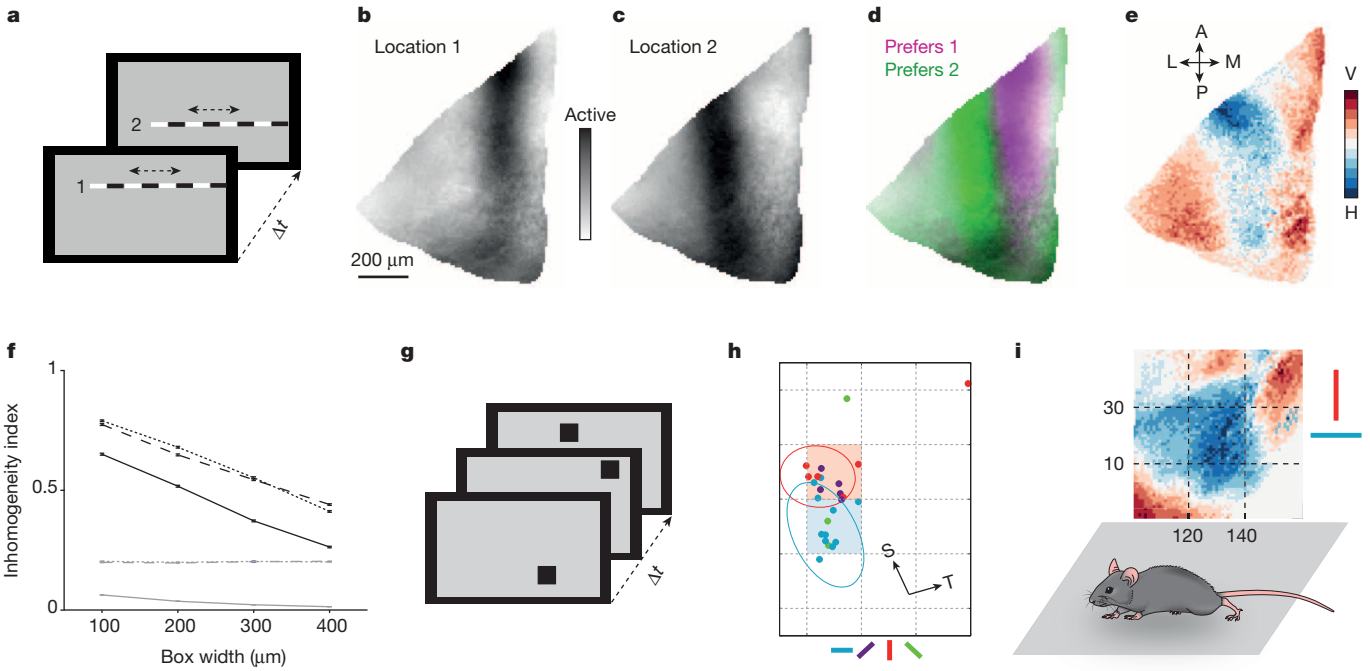


Figure 5 | Inhomogeneous coverage of contour orientation in the superior colliculus. **a**, Stimulus used to map retinotopy. **b–d**, Activation stripes elicited by the grating slit presented at positions 1 and 2 from **a**. Darker areas are more strongly activated. The reflectance change $\Delta R/R$ from black to white is 2×10^{-3} in both panels. **e**, Orientation map (from Fig. 4c). V, vertical; H, horizontal. **f**, Inhomogeneity indices for 3 mice from Fig. 4a–c for sampling windows 100–400 μm wide. Each black line corresponds to one mouse; each grey line corresponds to data from a shuffled orientation map. Error bars, s.e.m. **g**, Flashed spot stimulus used to map receptive fields in two-photon experiments. **h**, Receptive field centres (dots) for orientation-tuned cells in a field of view. Ellipses indicate two-dimensional Gaussian receptive field fits (radii 1 s.d.) for two representative cells. Image axes correspond to screen coordinates; dashed lines demarcate stimulus pixels. Red and blue squares indicate consensus receptive field centres of cells preferring vertical and horizontal bars, respectively. Arrows indicate projections of spherical coordinate axes from the animal's perspective. Arrow lengths, 5 degrees of visual angle. S, superior; T, temporal. **i**, Schematic of consequences of SC orientation columns for mouse vision. Image from Fig. 4a. In the temporal visual field, there are patches at high and low elevation where the SC prefers vertical bars and at intermediate elevation where it prefers horizontal bars. Approximate elevation and azimuth marked in degrees.

orientation patches spanned 200 μm to 400 μm in width (Fig. 5e; 3 animals), corresponding to ~30° of visual angle. Indeed, patches of the SC up to 300 μm wide were typically dominated by a particular preferred orientation (Fig. 5f). Thus the parcellation of orientation tuning is much coarser than the spatial projective field onto the SC. This implies that stimuli as large as 30° will be processed with some bias towards a certain contour orientation.

This unexpected result was scrutinized further. Intrinsic reflectance represents a bulk signal from the tissue, and the stimuli used to map orientation tuning and projective fields might excite two different sets of neurons. We therefore returned to two-photon calcium imaging to examine individual neurons. For each orientation-tuned neuron in several volumes we mapped the visual receptive field with small flashed spots (Fig. 5g). The receptive fields were comparable in size (median diameter, 10°; median area, 74°^{2}) to off-type receptive fields reported in anaesthetized mice[14]. Neurons were grouped by their preferred orientations into bins centred on 0, 45, 90 or 135°. The receptive field centres of cells preferring the same orientation were found clustered in visual space, separated from those preferring a different orientation (Fig. 5h). Moreover, there was little overlap between the receptive fields of cells preferring different orientations (Fig. 5h). Consistent with these observations, cells with a given orientation preference responded much more strongly to spots flashed in the consensus receptive field centre (Fig. 5h) of the cells sharing that preferred orientation (median 100%, mean 81%, fraction of peak response; 4 animals, 63 cells) than to the receptive field centre of cells with a different preferred orientation (median 29%, mean 38%; $P < 0.001$, Kruskal–Wallis test). Thus, both intrinsic imaging and two-photon calcium imaging reveal that adjacent orientation columns in the SC survey largely non-overlapping regions in space, in violation of position invariance and uniform coverage. These orientation patches cover large regions of the visual field (~30°) compared to the mouse's acuity (~2°)[20]. Furthermore, the absolute magnitude of this inhomogeneity is substantial. The average orientation-tuned neuron (OSI = 0.30) has a 2:1 bias in favour of the preferred orientation, so the relative gain for orthogonal orientations in adjacent patches may differ by a factor of 4. For regions of the visual field spanning several tens of degrees, the SC favours one orientation and responds less well to stimuli of other orientations (Fig. 5i).

It is not apparent how columnar architecture arises in the SC. Perhaps afferents from distinct retinal ganglion cell subtypes are routed to different orientation patches by different rules. Indeed, the terminal arbors of certain retinal ganglion cell types show a vertical columnar structure, albeit on a slightly finer scale[21,22]. Additional hints at non-uniform anatomy are the lattices formed deeper in the SC by cholinergic[23–25] and nigro-collicular[26] fibres, again on the scale of several hundred micrometres. The tuning field for contour orientation (Fig. 5i) may also relate to regional specializations in the SC. For example, stimulation of regions of the SC that survey the upper or lower visual field elicits avoidance or approach behaviours, respectively[27–29]. This subdivision seems coarser than the observed orientation tuning maps, and a more refined study of behaviours supported by the SC, along with further exploration of functional architecture in response to diverse stimuli[30] across the full retinotopic map, will help in understanding the functional significance of these columns in the SC.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 6 August; accepted 24 November 2014.
Published online 17 December 2014.

1. Masland, R. H. The neuronal organization of the retina. *Neuron* **76,** 266–280 (2012).
2. Dhande, O. S. & Huberman, A. D. Retinal ganglion cell maps in the brain: implications for visual processing. *Curr. Opin. Neurobiol.* **24,** 133–142 (2014).
3. Wässle, H. Parallel processing in the mammalian retina. *Nature Rev. Neurosci.* **5,** 747–757 (2004).
4. May, P. J. The mammalian superior colliculus: laminar structure and connections. *Prog. Brain Res.* **151,** 321–378 (2006).
5. Ohki, K., Chung, S., Ch'ng, Y. H., Kara, P. & Reid, R. C. Functional imaging with cellular resolution reveals precise micro-architecture in visual cortex. *Nature* **433,** 597–603 (2005).
6. Hubel, D. H. & Wiesel, T. N. Receptive fields and functional architecture of monkey striate cortex. *J. Physiol. (Lond.)* **195,** 215–243 (1968).
7. Blasdel, G. G. & Salama, G. Voltage-sensitive dyes reveal a modular organization in monkey striate cortex. *Nature* **321,** 579–585 (1986).
8. Grinvald, A., Lieke, E., Frostig, R. D., Gilbert, C. D. & Wiesel, T. N. Functional architecture of cortex revealed by optical imaging of intrinsic signals. *Nature* **324,** 361–364 (1986).
9. Harris, K. D. & Mrsic-Flogel, T. D. Cortical connectivity and sensory coding. *Nature* **503,** 51–58 (2013).
10. Chen, T. W. *et al.* Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499,** 295–300 (2013).
11. Dräger, U. C. & Hubel, D. H. Topography of visual and somatosensory projections to mouse superior colliculus. *J. Neurophysiol.* **39,** 91–101 (1976).
12. Dräger, U. C. & Hubel, D. H. Physiology of visual cells in mouse superior colliculus and correlation with somatosensory and auditory input. *Nature* **253,** 203–204 (1975).
13. Wang, L. *et al.* Direction-specific disruption of subcortical visual behavior and receptive fields in mice lacking the β2 subunit of nicotinic acetylcholine receptor. *J. Neurosci.* **29,** 12909–12918 (2009).
14. Wang, L., Sarnaik, R., Rangarajan, K., Liu, X. & Cang, J. Visual receptive field properties of neurons in the superficial superior colliculus of the mouse. *J. Neurosci.* **30,** 16573–16584 (2010).
15. Zhao, X., Chen, H., Liu, X. & Cang, J. Orientation-selective responses in the mouse lateral geniculate nucleus. *J. Neurosci.* **33,** 12751–12763 (2013).
16. Andermann, M. L., Kerlin, A. M., Roumis, D. K., Glickfeld, L. L. & Reid, R. C. Functional specialization of mouse higher visual cortical areas. *Neuron* **72,** 1025–1039 (2011).
17. Swindale, N. V., Shoham, D., Grinvald, A., Bonhoeffer, T. & Hubener, M. Visual cortex maps are optimized for uniform coverage. *Nature Neurosci.* **3,** 822–826 (2000).
18. Bosking, W. H., Crowley, J. C. & Fitzpatrick, D. Spatial coding of position and orientation in primary visual cortex. *Nature Neurosci.* **5,** 874–882 (2002).
19. Yu, H., Farley, B. J., Jin, D. Z. & Sur, M. The coordinated mapping of visual space and response features in visual cortex. *Neuron* **47,** 267–280 (2005).
20. Prusky, G. T. & Douglas, R. M. Characterization of mouse cortical spatial vision. *Vision Res.* **44,** 3411–3418 (2004).
21. Huberman, A. D. *et al.* Genetic identification of an On-Off direction-selective retinal ganglion cell subtype reveals a layer-specific subcortical map of posterior motion. *Neuron* **62,** 327–334 (2009).
22. Hong, Y. K., Kim, I. J. & Sanes, J. R. Stereotyped axonal arbors of retinal ganglion cell subsets in the mouse superior colliculus. *J. Comp. Neurol.* **519,** 1691–1711 (2011).
23. Ramon-Moliner, E. Acetylthiocholinesterase distribution in the brain stem of the cat. *Ergeb. Anat. Entwicklungsgesch.* **46,** 7–53 (1972).
24. Graybiel, A. M. A stereometric pattern of distribution of acetylthiocholinesterase in the deep layers of the superior colliculus. *Nature* **272,** 539–541 (1978).
25. Chevalier, G. & Mana, S. Honeycomb-like structure of the intermediate layers of the rat superior colliculus, with additional observations in several other mammals: AChE patterning. *J. Comp. Neurol.* **419,** 137–153 (2000).
26. Mana, S. & Chevalier, G. The fine organization of nigro-collicular channels with additional observations of their relationships with acetylcholinesterase in the rat. *Neuroscience* **106,** 357–374 (2001).
27. Dean, P., Redgrave, P., Sahibzada, N. & Tsuji, K. Head and body movements produced by electrical stimulation of superior colliculus in rats: effects of interruption of crossed tectoreticulospinal pathway. *Neuroscience* **19,** 367–380 (1986).
28. Sahibzada, N., Dean, P. & Redgrave, P. Movements resembling orientation or avoidance elicited by electrical stimulation of the superior colliculus in rats. *J. Neurosci.* **6,** 723–733 (1986).
29. Dean, P., Mitchell, I. J. & Redgrave, P. Responses resembling defensive behaviour produced by microinjection of glutamate into superior colliculus of rats. *Neuroscience* **24,** 501–510 (1988).
30. Basole, A., White, L. E. & Fitzpatrick, D. Mapping multiple features in the population response of visual cortex. *Nature* **423,** 986–990 (2003).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank E. Soucy and J. Greenwood for assistance with instrumentation; M. Joesch, A. Krishnaswamy, D. Kostadinov, S. Pashkovski, A. Giessel, T. Dunn, G. Keller, P. Kaifosh, M. Amoroso, and H. Asari for software; M. Andermann, V. Bonin, and F. Engert for advice on microscope design; J. Cohen for headplate designs; D. Anderson, K. Blum, B. Ölveckzy, and J. Sanes for critical reading of the manuscript; and J. Sanes for providing laboratory space and support to E.H.F. E.H.F. was supported by NIH T32 NS007484 and a Howard Hughes Medical Institute-Helen Hay Whitney Foundation fellowship. Additional support was provided by an NIH grant to M.M.

Author Contributions E.H.F. designed the study, performed all experiments, interpreted results, and wrote the manuscript. M.M. helped design the study, interpret results, and write the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.M. (meister@caltech.edu) or E.H.F. (evan_feinberg@post.harvard.edu).

METHODS

Mice. Experiments were conducted on adult C57BL6/J mice of both sexes (ages 2–8 months, Jackson labs). All procedures were performed in accordance with institutional guidelines. Mice were first anaesthetized with ketamine, xylazine, and acepromazine (60 mg per kg, 7.5 mg per kg, and 3 mg per kg, respectively) and placed in a stereotaxic device with eyes covered with ophthalmic ointment. A custom head plate (titanium, 1 mm thickness, eMachineShop) was bonded to the skull (ESPE Adper Scotchbond, 3M), roughly centred on lambda, parallel to the long axis of the mouse and at a pitch of 15 ± 5°. In some mice, viral injections were made through a small burr hole drilled over the rostral tip of the SC (injection coordinates 0.2–0.4 mm caudal to interaural line, 0.3–0.7 mm lateral). A glass pipette (25–35 μm tip) loaded with a 2:1 mixture of AAV2/1.hSyn1.GCaMP6s.WPRE.SV40 (Penn Vector Core) and 20% mannitol in saline was advanced into the tissue and a hydraulic injector was used to inject 90–270 nl over several minutes. The pipette was left in place for 3 min before progressing to the next depth. Injections were made at three depths, typically 1.3, 1.15, and 1 mm below lambda. The pipette was then slowly removed and the hole overlaid with Kwik-Cast silicone elastomer (World Precision Instruments). Animals were given buprenorphine and carprofen (0.1 mg per kg, 5 mg per kg, respectively) for 48 h post-operatively.

Previous intrinsic signal imaging studies of the SC entailed ablation of the overlying visual cortex (VC)[31,32]. This procedure destroys strong reciprocal connections between the SC and VC that likely influence response tuning, a potentially serious caveat. We noticed that the posteromedial SC is not obscured by cortex and might be accessible for imaging without removal of VC. This wedge of the SC lies beneath the confluence of the superior sagittal and transverse sinuses (Fig. 1a)[33]. Blood absorbs infrared and visible light, and these vessels form an opaque barrier that stymied initial imaging attempts. Ablation of the transverse sinus in humans with arteriovenous malformations typically causes minimal or no neurological symptoms[34], indicating that it is not essential for healthy brain function. Moreover, its location within the dura and consequent lack of physical attachment to the brain surface suggested the possibility of dislodging it without damaging the underlying SC. Several previous studies used 'plugs' of glass or transparent silicone to apply pressure normal to the brain surface to flatten the tissue and minimize motion artefacts[16,35]; we reasoned that triangular plugs could be used to apply pressure parallel to the brain surface, in a fashion analogous to a snowplow, to anteriorly displace the transverse sinuses (Fig. 1b, c and Extended Data Fig. 1a–c). With this approach, we were able to implant acute or chronic imaging windows and expose triangular patches of the SC typically 800–1,000 μm on a side, corresponding to ~15–25% of the surface area of the SC. We have successfully imaged the SC with this preparation from 30 min to 6 months after plug implantation.

Five days to 3 months after implantation of head plates, mice were given dexamethasone (2 mg per kg), anaesthetized with isoflurane, and immobilized via their head plates in a custom holder. In mice previously injected with AAV, a 2–3 mm craniotomy was made over the SC, inferior colliculus, and part of the cerebellum, and a large flap was opened in the dura with a 30-gauge needle. The tissue was kept moist under artificial cerebrospinal fluid (ACSF). ACSF was wicked from the craniotomy to leave only a thin film of liquid over the SC and a small drop of uncured Kwik-Sil was applied to the SC. A plug bonded to a 5 mm circular coverslip was mounted on a suction cup and positioned over the craniotomy. The plug was quickly advanced downward into the uncured drop of silicone and then anteriorly to displace the dura and transverse sinuses. Cyanoacrylate (Vetbond, 3M) was applied to bond the coverslip to the skull and headplate. After a few minutes, suction was released and the suction cup withdrawn. Black dental cement (Ortho-Jet, Lang Dental) was then applied over the cyanoacrylate on the skull, head plate, and edges of the cranial window and allowed to set for ~30 min. For mice that had not been injected with AAV, a similar craniotomy was performed over the SC as well as VC, virus was injected into the SC and VC as previously described, and a plug attached to an 8 mm coverslip was implanted as above.

To aspirate cortex, a craniotomy was performed over the SC and VC. Cortex was slowly aspirated and the transverse sinus was severed and reflected. Once bleeding had ceased, the craniotomy was filled with uncured Kwik-Sil and a coverslip was pressed in place above and bonded as previously described.

At least 3 days after implanting cranial windows, mice were habituated to handling and head-fixing for at least 3 days before experiments began; mice were given at least 7 days after injection to permit GCaMP expression, and indistinguishable results were obtained 7 days to 3 months after AAV injection. Mice were head-fixed on a 12 cm diameter circular treadmill (Ware Flying Saucer). The underside of the wheel was painted with alternating stripes of black and silver and illuminated with 940 nm light-emitting diodes (LEDs). A pair of photodiodes measured reflectance of the stripes and thus encoded wheel motion for steps >5 mm. Analogue signals were recorded and analysed in Matlab.

Plugs. Uncured Kwik-Sil (World Precision Instruments) was pressed between two ~2.5 cm square blocks of acrylic (previously sterilized with 70% ethanol) separated by 0.75 mm shim stock. The silicone was allowed to cure for at least 15 min and transferred to a sterile Petri dish. A scalpel was used to cut triangular prisms roughly 1 mm tall and 1.5 mm wide. Care was taken to avoid use of any portion of the silicone sheet containing bubbles or lint. Surfaces of the silicone plugs were cleaned with transparent adhesive tape to remove dust. A corona treater (Electrotechnic products) was used to activate the surfaces of a silicone plug and a coverslip (5 or 8 mm, number 1 thickness, Warner) and the plug was placed on the coverslip with the activated surfaces touching. Plugs were placed in a sterile Petri dish and bonding was allowed to proceed overnight in a hybridization oven at 60–70 °C. To implant plugs, small suction cups were fabricated by bevelling a 25-gauge needle to ~45° and mounting the needle with the aperture facing downward on a micromanipulator with a 20 ml syringe attached through flexible tubing. A small drop (~1 mm diameter) of ACSF was set on a clean block of acrylic and the tip of the syringe was placed in the drop. Kwik-sil was applied over the tip and allowed to set for at least 15 min before use.

Two-photon microscope. Two-photon imaging was performed on a custom-built microscope controlled by software written in Labview (National Instruments). A mode-locked Ti:sapphire laser (Mai-Tai DeepSee, Newport) with group delay dispersion compensation was scanned by galvanometers (Cambridge) through a 20× 1 NA water-immersion objective (Olympus). GCaMP6s was excited at 920 nm and laser power at the sample plane was typically 15–50 mW. Imaging in the SC was performed 50 μm to 300 μm below the surface, and imaging in layer 2/3 of VC was performed 150 μm to 300 μm below the surface. A 300 × 150 μm field of view was scanned at 8 Hz as a series of 300 × 150 pixel images. Emitted light was collected with a T600/200dcrb dichroic (Chroma) and a 610dxcr dichroic (Chroma) to split green and red light (no red fluorophore was used in this study); green light passed through a HQ600/200M-2P bandpass filter (Chroma) and was detected by a multialkali photomultiplier tube (R3896, Hamamatsu). Artefacts of the strobed stimulus were eliminated by discarding 10 pixels on either end of each line to yield 280 × 150 pixel images.

Intrinsic imaging microscope. Reflectance of a 735 nm LED (Thorlabs) was collected using a CCD camera (Flea3, Point Grey), through a 5× 0.14 NA air objective (Mitutoyo) used as a 2.5× objective with a short (f = 100 mm) tube lens. Images of 640 × 480 pixels at 8-bit resolution were acquired at 114 or 120 Hz and binned to 6 Hz, 160 × 120 pixels. Acquisition and analysis used custom software written in Labview and Matlab (Mathworks).

Visual stimuli. Stimuli were generated in Psychtoolbox3 (Matlab) and presented on an LCD screen (Dell, U2312HM) centred 23 cm away from the mouse's eye, angled at 20° in pitch and yaw to minimize fisheye distortion. Stimuli were presented on a square (1,080 × 1,080 pixel) region of the screen. Between experiments the monitor was maintained at a constant background grey level. To minimize interference of the stimulus with fluorescence detection, the monitor was strobed for 2 μs at the end of each scan line (1,200 Hz; luminance of grey screen ~1.25 cd m^{-2}, maximum brightness ~1/80 of unmodified monitor). Mice see red poorly, and all stimuli presented used only the green and blue channels of the monitor. The red channel was used to convey stimulus timing to synchronize with fluorescence acquisition; red bars flickered periodically at the bottom of the screen, which was covered with black tape. Drifting bars were 40 pixels wide (2–3°) and drifted at a speed of 240 pixels per s (12–18° s^{-1}). Flashed spots were presented as a 10 × 10 grid of 5–8° black squares. Each spot appeared for 500 ms and was followed by 500 ms of grey screen. All stimuli in two-photon experiments were presented in a pseudorandom sequence with interspersed blank periods (sampling with replacement) within each stimulus block (typically 6–8 blocks per experiment), with a different random seed for each block.

For intrinsic imaging, the same monitor was used. Due to the larger number of repetitions required, in some experiments the drifting bar stimuli were presented with interspersed blank frames omitted. To map the projective fields of slit gratings, a square wave grating with 100% contrast, spatial frequency of 0.5–0.8 cycles per degree (cpd), and temporal frequency of 1 Hz, switching direction after each cycle, was presented for 8 s through an aperture of the same dimensions as the drifting bars. To map the projective fields of grating patches, the same square grating was presented through a square aperture (10–15°) that alternated between two abutting locations every 8 s, changing to a randomly chosen (with replacement) cardinal direction every 1 s.

Calcium imaging analysis. Brain motion during imaging was corrected using TurboReg (ImageJ) or software written in Python[36]. Elliptical regions of interest (ROIs) were drawn manually in Matlab and fluorescence traces extracted and neuropil signals subtracted. Neuropil tended to share the orientation tuning of the embedded cells (Extended Data Fig. 5). Because much of the neuropil derives from processes of the local cells, if local cells are tuned alike, the surrounding neuropil will show similar tuning, as in orientation columns in cat visual cortex[5,37]. Thus, this observation was consistent with the single-cell data, but also presented a potential experimental confound, because out-of-focus fluorescence from neuropil leaks into

the signals recorded from individual cells[38,39] (Extended Data Fig. 5b), and contamination by a tuned neuropil signal could bias measurements of a cell's true preferred orientation. The true fluorescence signal of a neuron is $c = r - (f \times n)$, with r the raw fluorescence signal of the ROI containing the cell, f the out-of-focus neuropil contamination factor, and n the fluorescence signal of the surrounding neuropil. To estimate the extent of this contamination, small non-vertical blood vessels were identified, as in previous studies[39], and cell-sized 'holes' that likely corresponded to uninfected cells. The ratio of their brightness and that of the surrounding neuropil was measured, and this routinely yielded estimates of $f = \sim 0.5$, a value similar to that obtained by another group using an objective with NA 1 (ref. 39). As a result, all data presented in this study were processed using this value of f. A shell of radius 20 μm centred on each cell was taken from the masked image, excluding all cells, and used to calculate the corrected signal for neuropil contamination. Cells with bright, filled nuclei, known to have slower kinetics and/or aberrant responses[40], were excluded from analysis but included in masks for neuropil subtraction[10,40]. Nevertheless, the true value of f may vary slightly within a field of view, and to ensure that the central results were robust to this variation we repeated the analysis with values of f from 0.3 to 0.9. For each value of f in this range, neurons were much more similar to the surrounding neuropil and each other than chance (Extended Data Fig. 5c, d), confirming that cells are indeed tuned similarly to their neighbours and the surrounding neuropil.

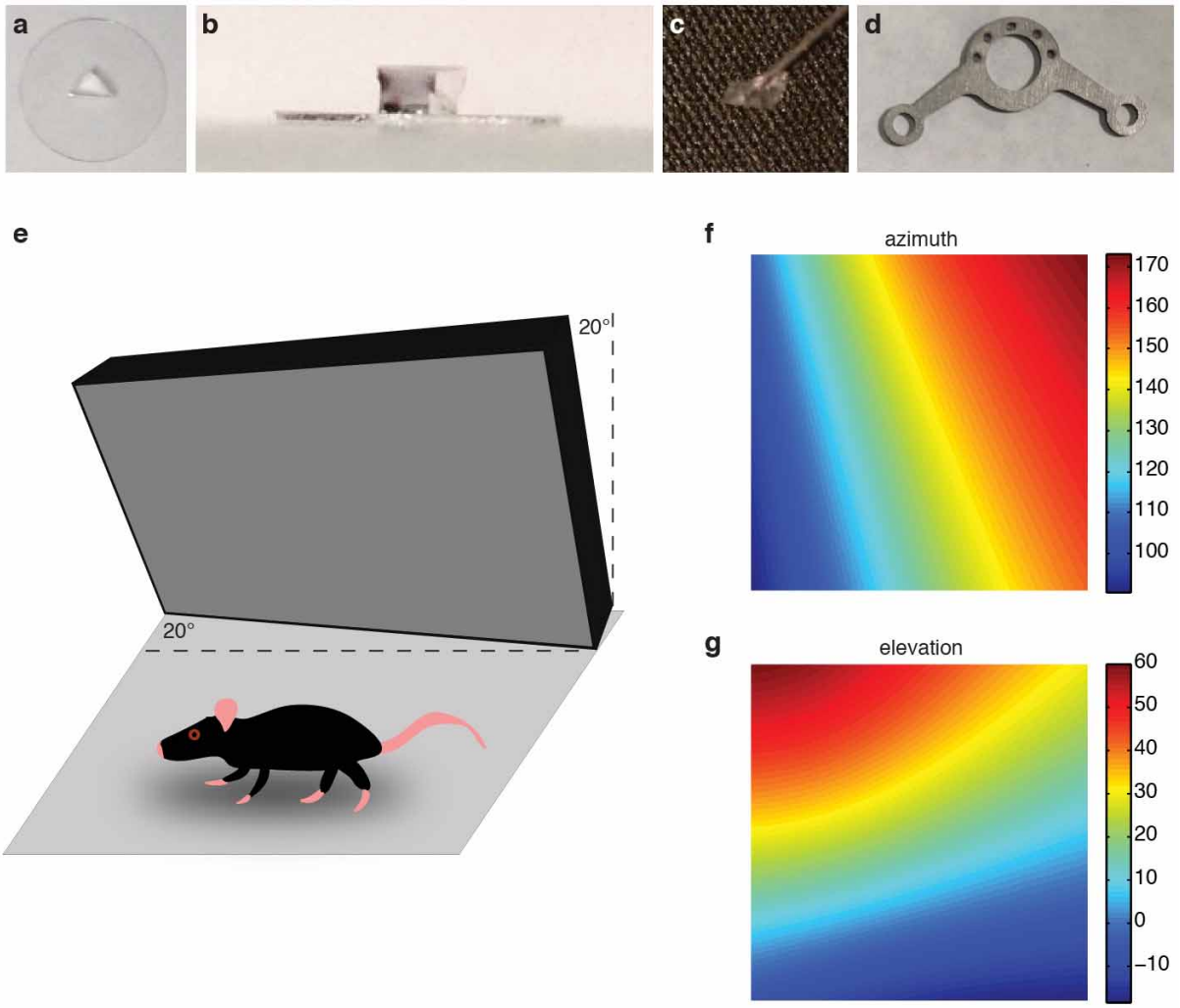
On occasion, frames with large-amplitude motion were not registered correctly. These frames can be easily identified because activity should cause only increases in the fluorescence signal for an ROI, while movement of a cell into or out of an ROI is likely to be associated with either increases or decreases in fluorescence[35]. To detect these events, we estimated the baseline fluorescence for each ROI as the mean of the lower half of its fluorescence intensity over the course of the movie, and discarded frames in which the fluorescence of any ROI was more than 3 s.d. below baseline. Slow baseline fluctuations were removed by subtracting the eighth percentile value from a moving window 15 s wide centred on each frame[35]; the mean value of the eighth percentile value for the entire trace was then added back to allow measurement of fold changes. The response to each direction ($\Delta F/F$) was measured as $(F/F_0)-1$, with F the instantaneous ROI intensity and F_0 the mean fluorescence intensity during stimulus blanks (grey screen). $\Delta F/F$ values for each presentation of a stimulus were averaged and the peak $\Delta F/F$ for each direction was used to compute orientation preference; similar results were obtained using the mean $\Delta F/F$ for each ROI. Responses to directions separated by 180° were summed to compute orientation preference. The orientation selectivity index (OSI) was calculated as $(R_{\mathrm{pref}} - R_{\mathrm{ortho}})/(R_{\mathrm{pref}} + R_{\mathrm{ortho}})$, with R_{pref} the $\Delta F/F$ to the orientation eliciting the strongest response and R_{ortho} the $\Delta F/F$ to the orthogonal orientation[15]. Cells with OSI ≥ 0.15 (~4:3 preferred:null response) were classified as orientation-selective. The preferred orientation was defined as the weighted vector sum over the range of presented stimulus angles.

To map receptive fields with flashed spots, stimuli were presented in rapid succession, necessitating a faster baseline filter (3 s) to compensate for the slow decay kinetics of the GCaMP6s. The consensus centre square for each field of view, averaging all cell bodies and neuropil, was identified and the mean responses to spots on the periphery, well outside the receptive fields, and blanks were averaged to compute baseline fluorescence F_0 for each ROI. The response $\Delta F/F$ was computed as $(F/F_0)-1$, with F representing the peak of the averaged response of all presentations for each spot location. To compare receptive field overlap, preferred orientation for each cell was binned to 0, 45, 90, or 135°, according to the presented orientation eliciting the strongest response. Next, the number of cells preferring each orientation was determined, and the orientations preferred by the largest fraction of cells (orientation 1) and second-largest fraction of cells (orientation 2) were analysed. The peak square location for each cell preferring orientation 1 or orientation 2 was defined as the location that elicited the maximal responses from the plurality of cells sharing each preferred orientation. A two-dimensional Gaussian was fit to each orientation-tuned cell's responses to the flashed spots and the receptive field size was set as the area of an ellipse of radius of one s.d., using a spot area of 7.5°.

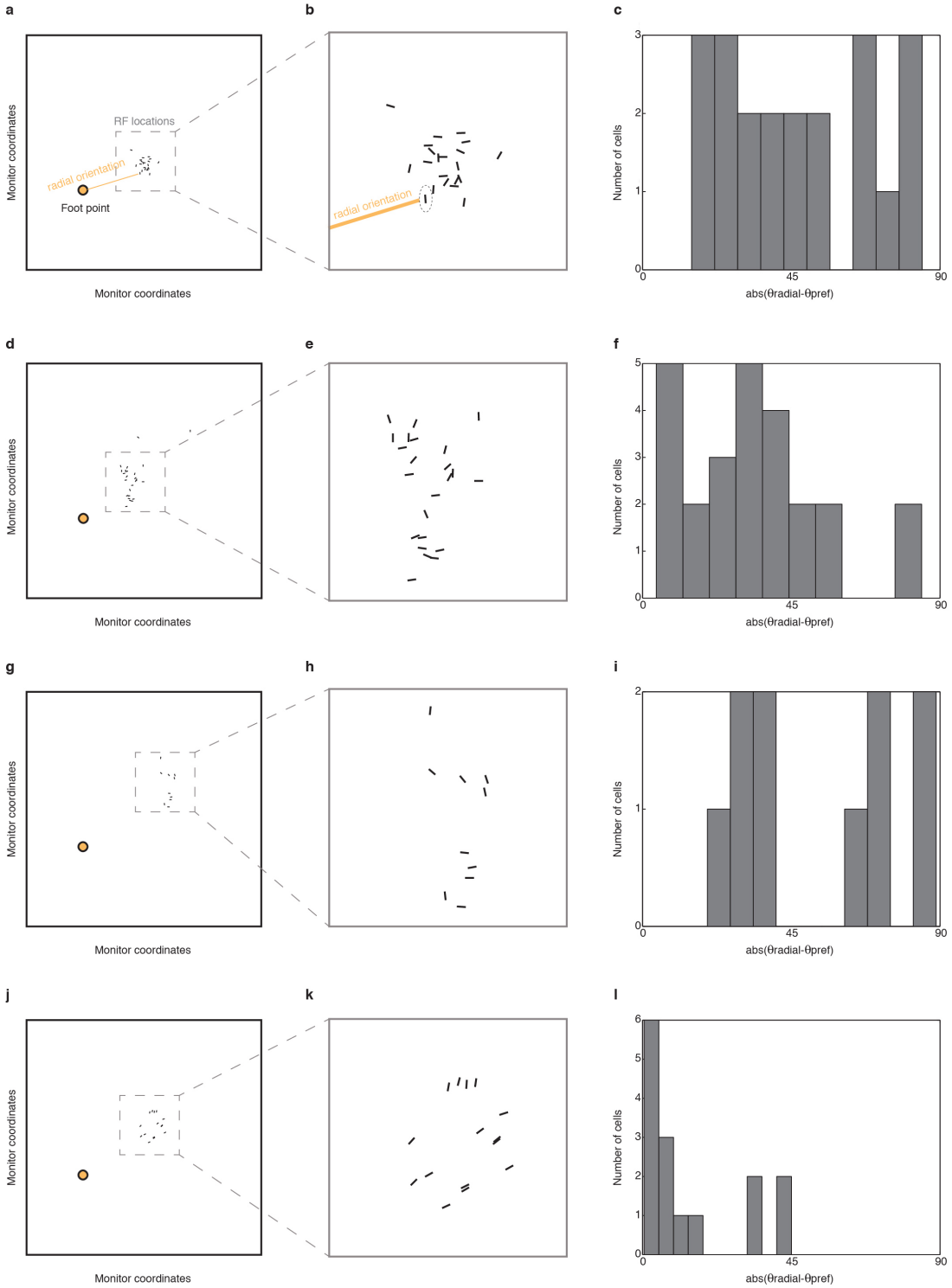
Intrinsic imaging analysis. No temporal or spatial filtering of intrinsic imaging data was performed. To map orientation columns, mean reflectance changes while bars drifted in both directions were summed for each cardinal axis. An ROI was drawn manually over the SC to exclude signals from the large surrounding blood vessels, and the ratio of responses to the vertical and horizontal axes was determined. To measure the projective fields of slit gratings, the ratio of the mean reflectance to a given grating position was taken to the mean reflectance when orthogonal bars were presented across the SC. To map projective fields with grating patches, the ratio of the mean reflectance when the stimulus was at either of the two locations was determined. Measurements of projective field peak-to-peak distances were made in ImageJ by drawing a rectangular ROI over each projective field, roughly orthogonal to the bar, and measuring the averaged line profile over the ROI. These line profiles were smoothed and full-width at half-maximum was measured in Matlab. To measure inhomogeneity indices, pixels were classified as preferring horizontal or vertical if they were in the upper 40% of responses to either orientation, with the remaining pixels classified as untuned. A window of indicated size was rastered across the orientation map. At any position in which more than half of the pixels were within the SC mask, the inhomogeneity index was measured as, $abs\ (h - v)/(h + v + u)$ with h and v the number of pixels preferring horizontal and vertical bars, respectively, and u the number of untuned pixels. To generate the image in Fig. 5i, lines were manually fit to the projections of slit gratings (as in Fig. 5a–c) on the surface of the SC. The resulting grid was then aligned to a grid of the positions of the stimuli in the visual field using a thin plate spline in Matlab, and the same transform applied to the orientation map from that animal.

Statistical methods. No statistical method was used to predetermine sample size. Statistical comparisons were performed as Kruskal–Wallis tests, a non-parametric test that extends the Mann–Whitney U-test to multiple groups, or as Monte Carlo simulations. These tests do not assume normality of the data; nevertheless, all comparisons described yielded similar P values with statistical methods that assume normality (for example, one-way ANOVA).

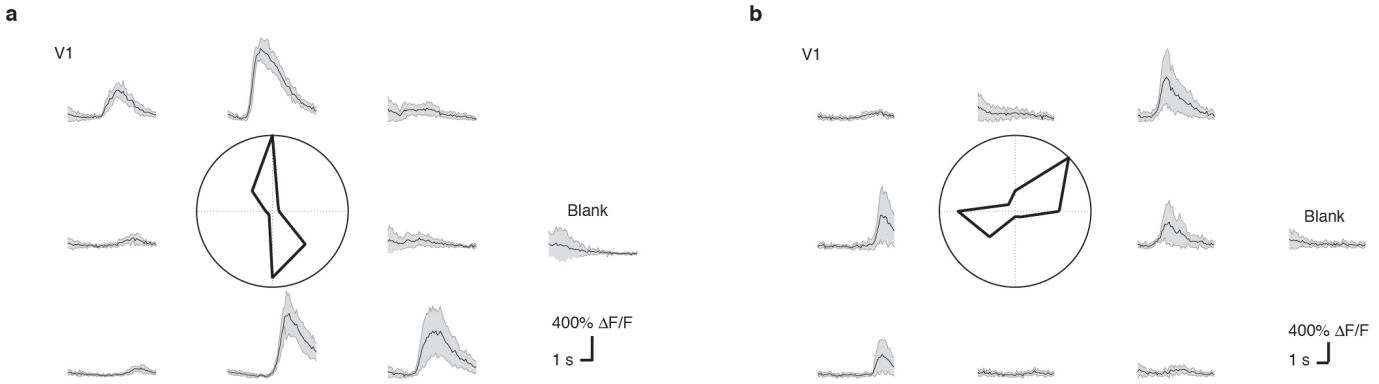
31. Mrsic-Flogel, T. D. *et al.* Altered map of visual space in the superior colliculus of mice lacking early retinal waves. *J. Neurosci.* **25,** 6921–6928 (2005).
32. Cang, J., Wang, L., Stryker, M. P. & Feldheim, D. A. Roles of ephrin-as and structured activity in the development of functional maps in the superior colliculus. *J. Neurosci.* **28,** 11015–11023 (2008).
33. Dorr, A., Sled, J. G. & Kabani, N. Three-dimensional cerebral vasculature of the CBA mouse brain: a magnetic resonance imaging and micro computed tomography study. *Neuroimage* **35,** 1409–1423 (2007).
34. Wong, G. K., Poon, W. S., Yu, S. C. & Zhu, C. X. Transvenous embolization for dural transverse sinus fistulas with occluded sigmoid sinus. *Acta Neurochirurgica* **149,** 929–935 (2007).
35. Dombeck, D. A., Khabbaz, A. N., Collman, F., Adelman, T. L. & Tank, D. W. Imaging large-scale neural activity with cellular resolution in awake, mobile mice. *Neuron* **56,** 43–57 (2007).
36. Kaifosh, P., Lovett-Barron, M., Turi, G. F., Reardon, T. R. & Losonczy, A. Septo-hippocampal GABAergic signaling across multiple modalities in awake mice. *Nature Neurosci.* **16,** 1182–1184 (2013).
37. Ohki, K. & Reid, R. C. *In vivo* two-photon calcium imaging in the visual system. *Cold Spring Harb. Protoc.* **2014,** 402–416 (2014).
38. Göbel, W. & Helmchen, F. *In vivo* calcium imaging of neural network function. *Physiology* **22,** 358–365 (2007).
39. Kerlin, A. M., Andermann, M. L., Berezovskii, V. K. & Reid, R. C. Broadly tuned response properties of diverse inhibitory neuron subtypes in mouse visual cortex. *Neuron* **67,** 858–871 (2010).
40. Tian, L. *et al.* Imaging neural activity in worms, flies and mice with improved GCaMP calcium indicators. *Nature Methods* **6,** 875–881 (2009).



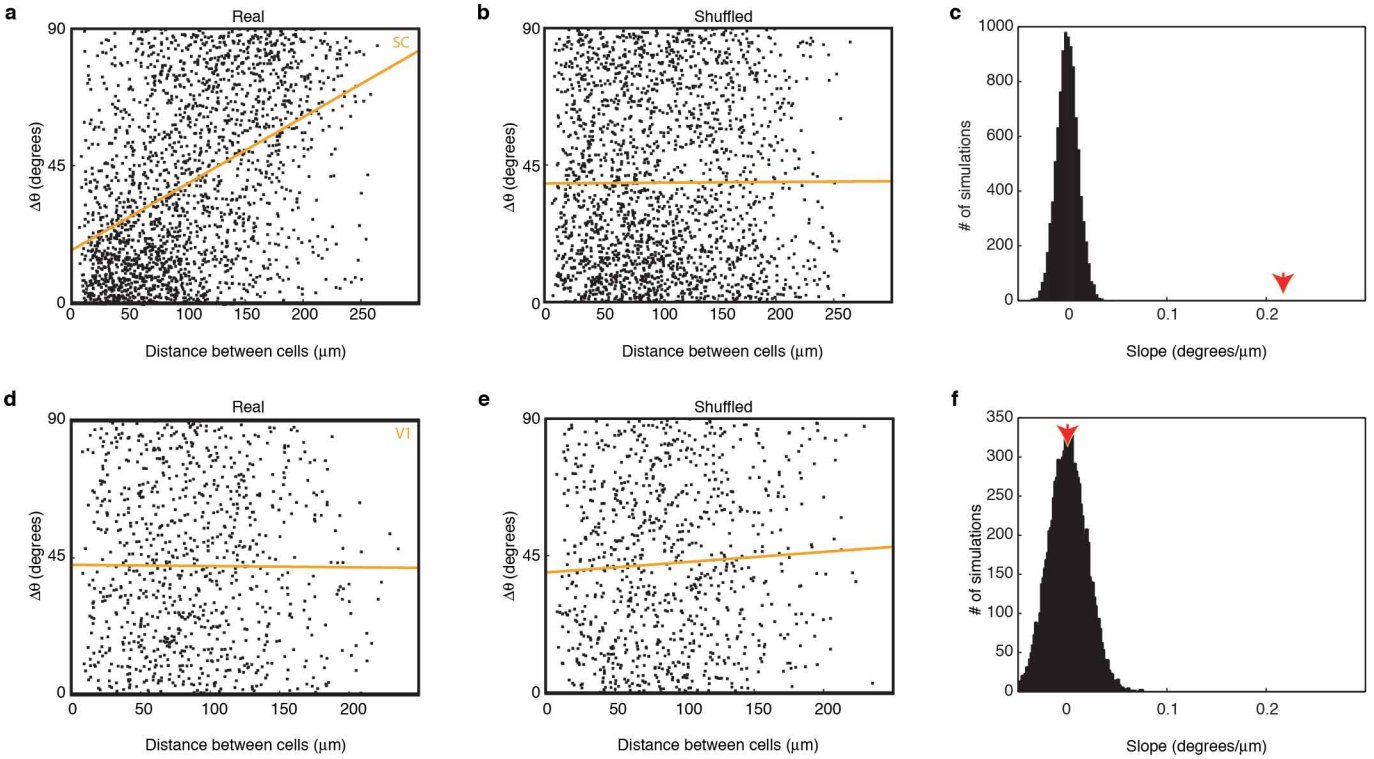
Extended Data Figure 1 | Plugs and head plates used for this study. **a**, Top view of triangular silicone plug attached to a 5 mm coverslip. **b**, Side view of the plug from **a**. **c**, Suction cup used to position plug. **d**, Standard headplate with 8 mm aperture. **e**, Monitor tilt used to reduce fisheye distortion. Perspective is exaggerated for clarity. The monitor was tilted such that it was 20° from vertical, with the top nearer the animal, and 20° from the animal's anterior–posterior axis, with the right edge closer to the animal. Note that stimuli were presented in a square area on the right side of the rectangular monitor. **f**, **g**, Azimuth and elevation in degrees for each pixel within the inscribed square area of the monitor on which stimuli were displayed. Values are given with respect to standard stereotaxic coordinates; because headplates were implanted at ~15° angles with respect to this plane, bars on screen are tilted. Curvature of elevation bars reflects the fact that iso-elevation lines are curved, not straight, much like latitude lines on a globe.



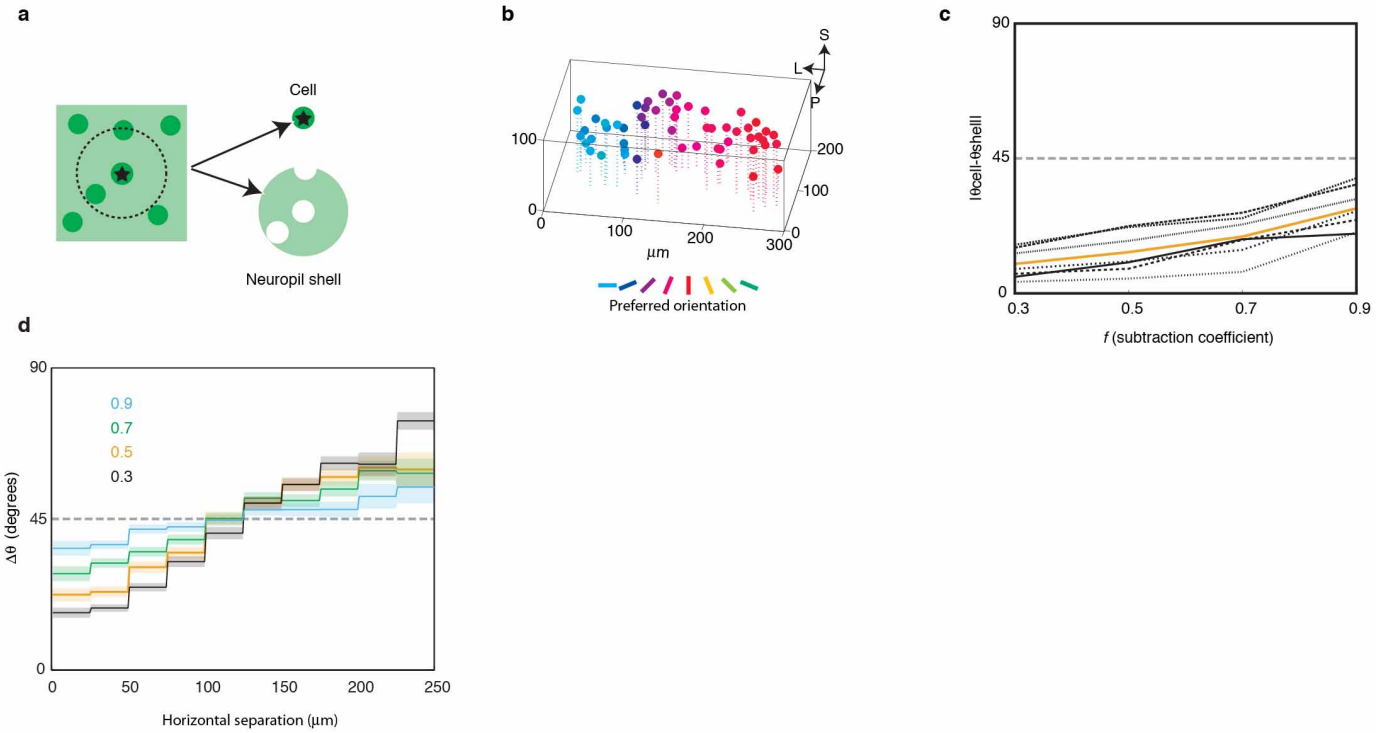
Extended Data Figure 2 | Orientation tuning does not reflect distortion effects of flat screen. **a**, Plot of foot point of monitor (FP), the point at which a line from the eye is perpendicular to the tangent screen, and all mapped receptive field centres for orientation-tuned SC neurons in one animal. Each line segment is positioned at a cell's receptive field centre and angled to reflect its preferred orientation. Orange line indicates the radial orientation relative to the FP for an example cell. Fisheye distortion will cause bars along the radial orientation relative to the foot point to appear relatively wider and faster than orthogonal bars along the tangential orientation, potentially biasing responses towards or away from the radial orientation. **b**, Enlarged view of inset area in **a** to show orientations more clearly. Note sharp transition in preferred orientation from bottom to top of panel relative to difference in radial orientation. Also note that the preferred orientation and the radial orientation vary with opposite handedness. **c**, Difference between radial orientation and preferred orientation for all cells in the plot. If the orientation map were due to fisheye distortion, preferred orientations should be similar to the radial orientation and the distribution should be centred at 0. Note that this distribution is biased away from 0 and centred between 45 and 90°. **d**–**f**, As in **a**–**c** for another animal. Note sharp transition in preferred orientations as in **b**, but with opposite handedness, and centring of distribution of preferred orientations between 0 and 45°. **g**–**i**, As in **a**–**c** for another animal. Note sharp transition in preferred orientations and bias of cells to orientations orthogonal to radial orientation. **j**–**l**, As in **a**–**c** for another animal. Note sharp transition in this field and a group of cells whose preferred orientations are close to radial.



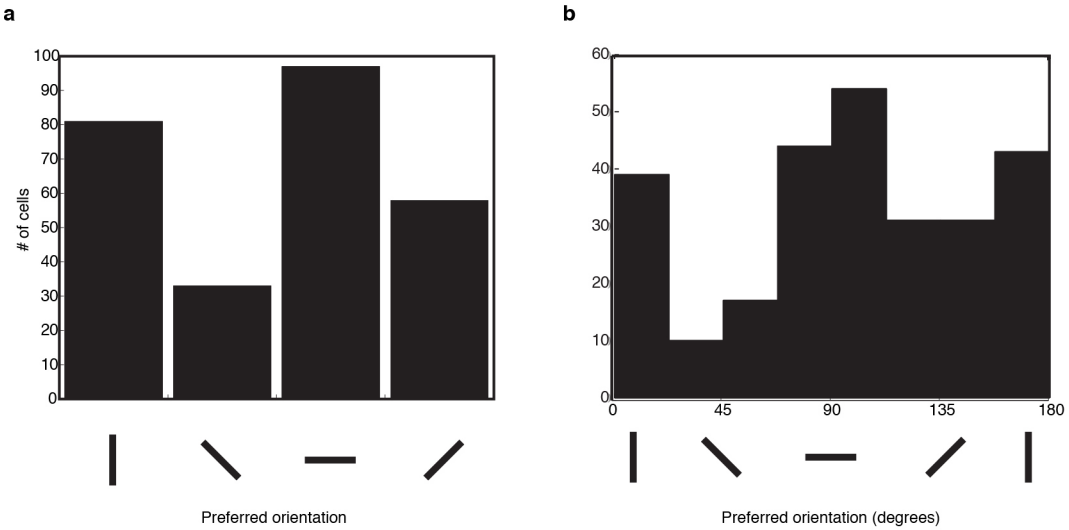
Extended Data Figure 3 | Sample responses of neurons in V1. a, **b**, Average $\Delta F/F \pm$ s.d. of two V1 neurons to 7 repetitions each of 8 directions of bar motion and a blank screen. Insets are polar plots for each cell.



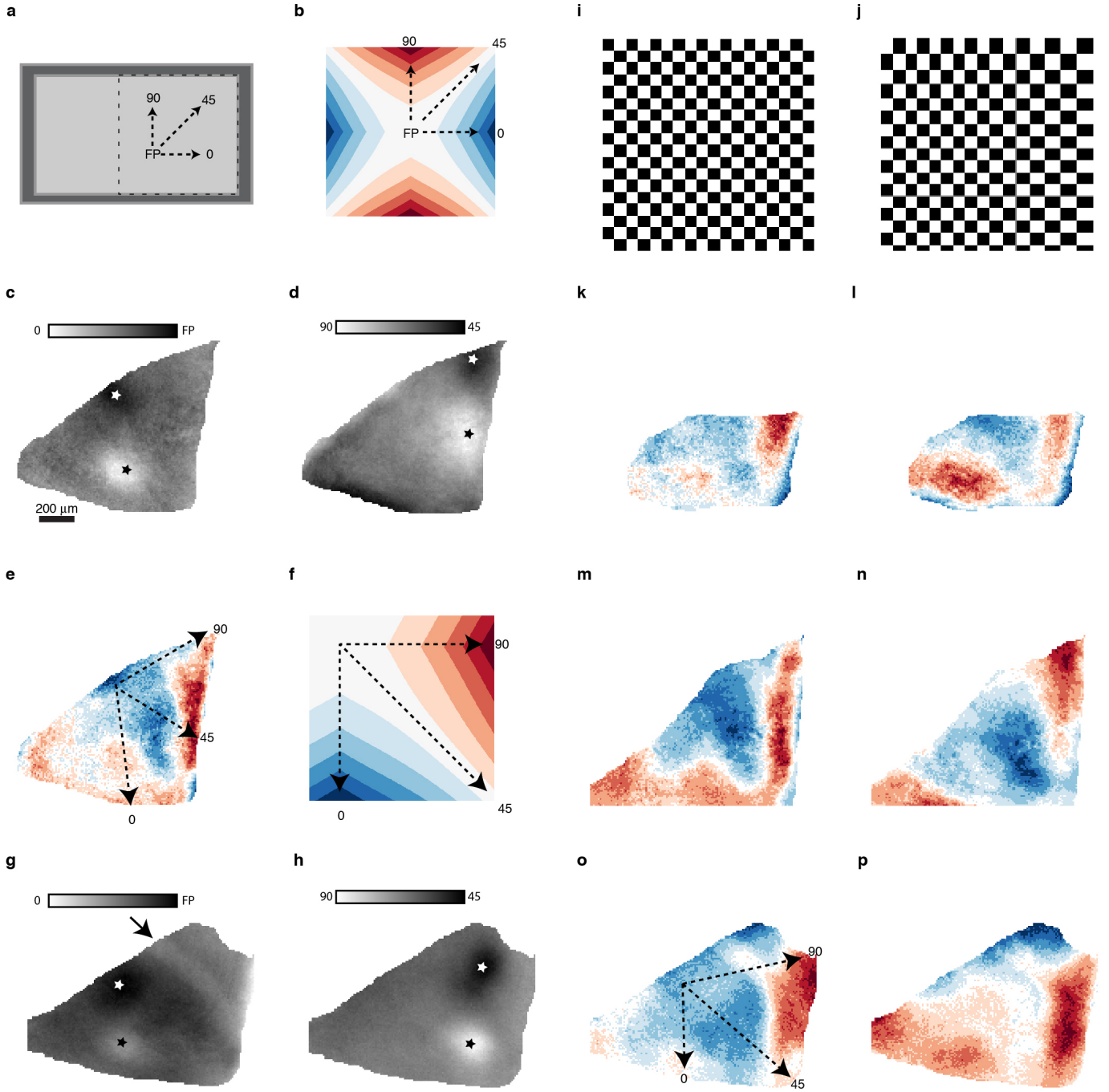
Extended Data Figure 4 | Monte Carlo simulation reveals significance of observed local similarity in the SC. **a**, Absolute value of the difference in preferred orientations plotted against horizontal distance in the SC. Orange line indicates linear fit to the difference in preferred orientation as a function of horizontal separation, yielding a line of best fit with a slope of $+22^\circ$ per 100 μm. **b**, As in **a** after shuffling all cell positions. A total of 10^5 independent shuffles were performed; shown are results from the final shuffle. This yielded a distribution of slopes with a mean $\pm$ s.d. of ($0 \pm 1^\circ$) per 100 μm. **c**, Histogram of slopes of best-fit lines from 10^5 independent Monte Carlo simulations. Arrowhead indicates slope from **a**. **d**–**f**, As in **a**–**c** for data from V1.



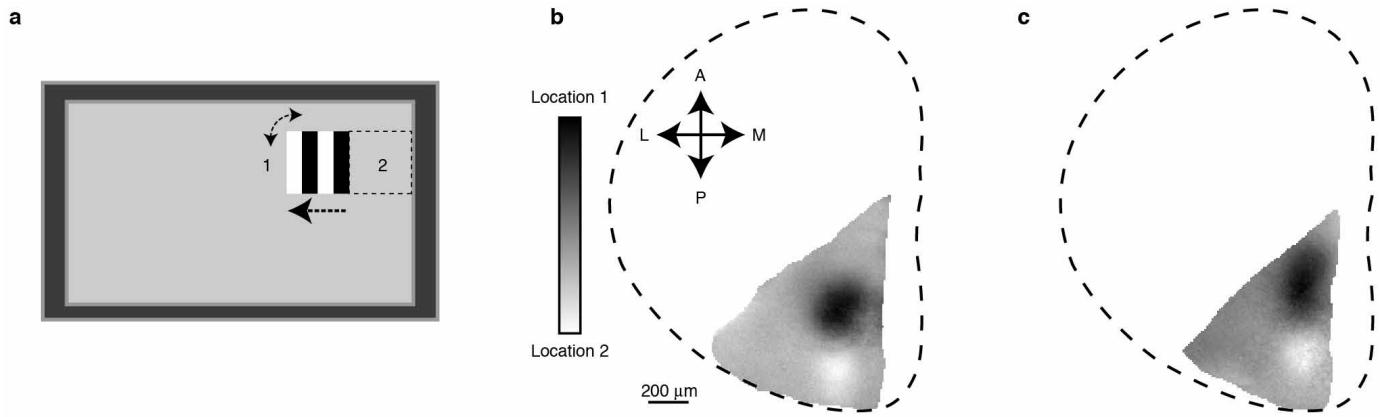
Extended Data Figure 5 | Orientation tuning of neuropil. **a**, Schematic of cells and surrounding neuropil shell. Signals are extracted from each cell and from the neuropil within 20 μm. The corrected signal of a cell c within an ROI r is $c = r - (f \times n)$, with n the signal of the neuropil shell and f the fractional contamination by out-of-focus neuropil. **b**, Orientation preferences for neuropil shells of orientation-tuned cells in Fig. 2a. **c**, Difference in preferred orientation of neurons and their neuropil shells over a range of values of neuropil subtraction coefficient f. Dashed horizontal line indicates chance. Each black line reflects median values from a single image volume; orange line is median value for all cells from 7 volumes. **d**, Effects are robust over a range of f values. Plotted are mean differences in preferred orientations against distance ± s.e.m as in Fig. 2d, from which the orange trace is reproduced. Because the neuropil is also sharply tuned, using high values of f will reduce the apparent similarity of neighbouring cells' orientation preferences. Nonetheless the similarity remains significant even at the excessively high f value of 0.9.



Extended Data Figure 6 | Distribution of preferred orientations in the SC. **a**, Preferred orientations of cells in the SC according to the presented orientation eliciting the strongest response. **b**, Preferred orientations of cells in the SC calculated from vector sums of responses to all orientations.



Extended Data Figure 7 | Orientation maps do not reflect fisheye distortion. **a**, Grating patches were presented at the foot point (FP), the point at which a line from the eye is perpendicular to the tangent screen, and at locations on the screen displaced from the FP along radial orientations of 0, 45, and 90°. **b**, Predicted map if fisheye distortion caused orientation tuning to be biased towards the radial orientation with respect to the FP. **c**, Projective field of foot point (black, indicated with white star) and patch at 0° (directly lateral on screen) from foot point (white patch, black star). **d**, As in **c** for patches at 90 and 45° with respect to foot point. **e**, Orientation map for this animal. Blue areas prefer horizontal bars, red areas prefer vertical bars, and arrows indicate lines from projective field of foot point to projective fields of patches. **f**, Expected orientation map according to distortion hypothesis. Note that the area at the projective field of the foot point should be untuned, and a line from the projective fields of the FP and a spot located at 0° relative elevation should pass from untuned areas to progressively more horizontal-preferring areas. Instead, it passes from a horizontal-preferring area at the FP to vertical-preferring areas as it moves to greater eccentricity. Trajectories along other projections of radial orientations (45 and 90°) are similarly poor fits to prediction. **g**, **h**, As in **b** and **c** for another animal. Orientation map for this animal in **o** also does not match prediction of fisheye distortion hypothesis. Arrow indicates shadow of blood vessel. **i**, **j**, Checkerboard pattern before and after 'pre-distortion' to offset fisheye effect. This pre-distortion was applied to change bar width by 1/cosine(θ), with θ the eccentricity from the FP, for both vertical and horizontal bar stimuli. **k**, **l**, Orientation maps for an animal in response to standard (**k**) and 'pre-distorted' (**l**) bar stimuli. In this animal the transverse sinus was not fully retracted and partially obscures the field of view. Note similarity of patterns in **k** and **l**. **m**, **n**, As in **k**, **l**, for the animal in **c**–**e** imaged on a different day. Comparison of maps in **e** and **m** reveals inter-trial variability, which is comparable to variability between standard and pre-distorted stimuli (**m** and **n**). **o**, **p**, As in **k**, **l** for a third animal. Map in **o** is overlaid with projective fields of points in visual field as in **e**. The reflectance change $\Delta R/R$ from black to white is 12×10^{-4} (**c**), 19×10^{-4} (**d**), and 18×10^{-4} (**g**, **h**). The reflectance change $\Delta R/R$ from red to blue is 4×10^{-4} (**e**, **k**, **o**), 5×10^{-4} (**l**, **p**), 7×10^{-4} (**m**), and 9×10^{-4} (**n**).



Extended Data Figure 8 | Alternative mapping stimulus reveals similar projective fields. **a**, Stimulus. Square grating patches alternate between two adjacent locations every 8 s. At each location the grating switches orientation randomly at 1 Hz. **b**, Map of responses elicited in animal from Fig. 4a. Responses to the two gratings span small patches on the surface of the SC. The reflectance change $\Delta R/R$ from black to white is 2×10^{-3}. **c**, As in **b** for animal from Fig. 4c. The reflectance change $\Delta R/R$ from black to white is 2×10^{-3}.

LETTER

doi:10.1038/nature14024

Mechanosensory interactions drive collective behaviour in *Drosophila*

Pavan Ramdya[1,2], Pawel Lichocki[2,3]†, Steeve Cruchet[1], Lukas Frisch[4], Winnie Tse[4], Dario Floreano[2] & Richard Benton[1]

Collective behaviour enhances environmental sensing and decision-making in groups of animals[1,2]. Experimental and theoretical investigations of schooling fish, flocking birds and human crowds have demonstrated that simple interactions between individuals can explain emergent group dynamics[3,4]. These findings indicate the existence of neural circuits that support distributed behaviours, but the molecular and cellular identities of relevant sensory pathways are unknown. Here we show that *Drosophila melanogaster* exhibits collective responses to an aversive odour: individual flies weakly avoid the stimulus, but groups show enhanced escape reactions. Using high-resolution behavioural tracking, computational simulations, genetic perturbations, neural silencing and optogenetic activation we demonstrate that this collective odour avoidance arises from cascades of appendage touch interactions between pairs of flies. Inter-fly touch sensing and collective behaviour require the activity of distal leg mechanosensory sensilla neurons and the mechanosensory channel NOMPC[5,6]. Remarkably, through these inter-fly encounters, wild-type flies can elicit avoidance behaviour in mutant animals that cannot sense the odour—a basic form of communication. Our data highlight the unexpected importance of social context in the sensory responses of a solitary species and open the door to a neural-circuit-level understanding of collective behaviour in animal groups.

Drosophila melanogaster is classified as a solitary species[7] but flies aggregate at high densities (>1 fly per cm^2) to feed[8] (Extended Data Fig. 1a, b and Supplementary Video 1), providing opportunities for collective interactions. Although groups affect circadian rhythms[9] and dispersal[10] in *Drosophila*, how social context influences individual sensory behaviours is unknown. To study this question, we developed an automated behavioural assay to track responses of freely-walking flies to laminar flow of air or an aversive odorant, 5% carbon dioxide (CO_2)[11,12]. Odour was presented to one half of a planar arena for 2 min (Fig. 1a and Extended Data Fig. 1c, d). Avoidance behaviour was quantified as the percentage of time a fly spent in the air zone during the second minute of a trial (Fig. 1b, c). Unexpectedly, isolated flies spent very little time avoiding this odour (Fig. 1d), despite the aversion to CO_2 observed in other assays[11,12]. However, increasing the number of flies was associated with substantial increases in odour avoidance (Fig. 1d and Extended Data Fig. 1e). This effect peaked at 1.13 flies per cm^2, a density typical for fly aggregates (Extended Data Fig. 1b) and was only apparent for flies in the odour zone (Fig. 1e and Extended Data Fig. 1f). Time-course analysis revealed that, within only a few seconds after odour onset, a larger proportion of flies in high-density groups had left the odour zone compared to isolated individuals (Fig. 1f; comparing 0.06 against 1.13 flies per cm^2, $P < 0.05$ for a Mann–Whitney *U*-test from 0.6 s onwards). Additionally, the motion of flies after odour onset was coherent at higher densities, with flies moving in the same direction, out of the odour zone; this effect was not observed for flies in the air zone (Extended Data Fig. 1g, h).

To determine the basis of these global behavioural differences, we examined the locomotion of individual flies. Single animals are typically sedentary but walk more when exposed to CO_2 (Extended Data Fig. 2a, b). In groups, however, we discovered that 63% of the time, the first walking response of a fly after odour onset coincided with proximity to a neighbouring fly (an 'Encounter': distance to a neighbouring fly < 25% body length; Fig. 2a–c and Supplementary Video 2). These Encounters were more frequent with increasing group density (Fig. 2d). Moreover, walking bouts (velocity > 1 mm s^{-1}) initiated during an Encounter ('Encounter Responses') were significantly longer than those spontaneously initiated in isolation (Fig. 2e). These observations indicated that inter-fly interactions might contribute to the enhanced odour avoidance of groups of flies.

We examined this possibility initially by computational simulation of the olfactory assay. The dynamics of our simulation were driven by three phenomena observed in behavioural assays (Fig. 2f). First, flies initiate more spontaneous bouts of walking in odour than in air (Extended Data Fig. 2a, b). Second, flies are more likely to turn and retreat after entering the odour zone from the air zone (Extended Data Fig. 2c). Third, close proximity to another fly elicits Encounter Responses in stationary flies (Fig. 2e and Extended Data Fig. 2d). Importantly, these elements could reproduce collective behaviour: higher numbers of simulated flies exhibited greater avoidance (Fig. 2g). While changing the olfactory parameters preserved stronger responses in groups than isolated individuals (Extended Data Fig. 2e–h), diminishing the Encounter Response probability could abolish and even reverse collective behaviour (Fig. 2h). These results suggested that Encounter Responses are a crucial component of *Drosophila* group dynamics.

To experimentally test the role of inter-fly interactions in collective behaviour, we sought to explain the mechanistic basis of Encounter Responses. Although our olfactory experiments were performed in the dark (Fig. 3a), the presence of light did not diminish Encounter Response frequency (Fig. 3a). Volatile chemicals are known modulators of many social behaviours[13,14], but putative anosmic flies (lacking known olfactory co-receptors) did not reduce Encounter Responses (Fig. 3a). By contrast, disruption of the mechanosensory channel NOMPC[5,6] significantly diminished Encounter Response frequency (Fig. 3a). These data suggested that mechanosensing is required for Encounter Responses.

By observing groups of flies at high spatiotemporal resolution, we found that active flies elicited motion in stationary animals through gentle touch of peripheral appendages (legs and wings; Fig. 3b and Supplementary Video 3). Leg touches took place exclusively on distal segments (Fig. 3b, inset) and resulted in spatially stereotyped walking reactions (Fig. 3c). These reactions were kinematically indistinguishable from Encounter Responses (compare Extended Data Fig. 3c and e; two-sample Kolmogorov–Smirnov test, $P = 0.07$; see Methods). This analysis indicates that appendage touch is the stimulus that elicits Encounter Responses. The precise stereotypy of these locomotor responses, similar to cockroach escape reactions[15], implies their dependence upon somatotopic neural circuits linking touch with movement.

As fly appendages also house taste receptors[16], we tested whether mechanical stimulation was sufficient to elicit Encounter Responses by tracking stationary flies following touch of appendages with a metallic

[1]Center for Integrative Genomics, Faculty of Biology and Medicine, University of Lausanne, Lausanne CH-1015, Switzerland. [2]Laboratory of Intelligent Systems, Institute of Microengineering, École Polytechnique Fédérale de Lausanne, Lausanne CH-1015, Switzerland. [3]Department of Ecology and Evolution, University of Lausanne, Lausanne CH-1015, Switzerland. [4]Master's Program in Microengineering, Institute of Microengineering, École Polytechnique Fédérale de Lausanne, Lausanne CH-1015, Switzerland. †Present address: Google Inc. (P.L.), 8 Rue de Londres, 75009 Paris, France.

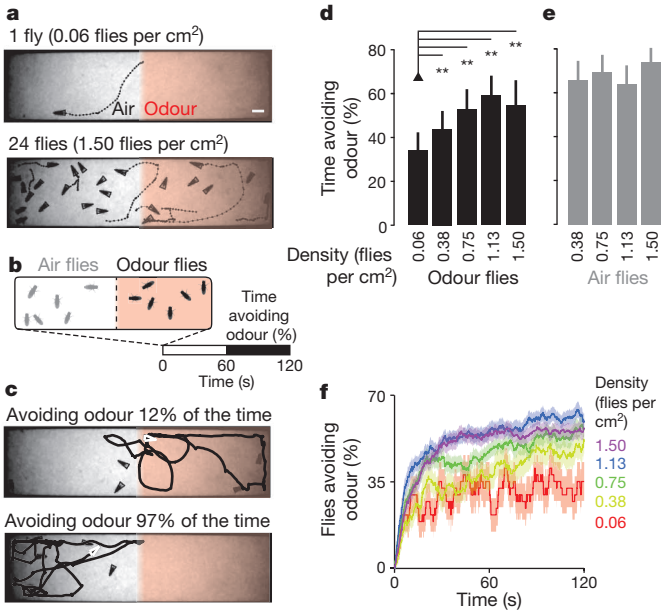


Figure 1 | Collective odour avoidance in *Drosophila*. **a**, Image of flies (triangles) and their trajectories (dashed lines) during 2 s in a two-choice olfactory assay. 5% CO_2 ('Odour') flows through the right half while air flows through the left half. Two densities of flies are shown (0.06 and 1.50 flies per cm^2). The scale bar is 2.5 mm. **b**, Schematic of the odour avoidance experiment. Flies in the odour zone at stimulus onset ($t = 0$) are measured for the time spent in the non-odour zone during the second minute of the experiment ('Time avoiding odour (%)'). **c**, Flies (white triangles) with a low (top) or high (bottom) per cent time avoiding odour. **d**, **e**, The per cent time avoiding the odour (mean and s.d.) for five different densities of flies starting in the odour zone (black bars) (**d**) and four densities of flies starting in the air zone (grey bars) (**e**). $n = 37, 38, 36, 35$, and 38 experiments for 0.06, 0.38, 0.75, 1.13, and 1.5 flies per cm^2 respectively. In this and all subsequent figures, unless otherwise stated, a single asterisk (*) denotes $P < 0.05$ and a double asterisk (**) denotes $P < 0.01$ for a Bonferroni-corrected paired sample *t*-test (bar plot comparisons) or a Mann–Whitney *U*-test (boxplot comparisons). **f**, The proportion of flies outside of the odour zone over the entire experiment. The mean (solid line) and s.e.m. (transparency) are colour-coded for each density (*n* is as for panels **d**, **e**).

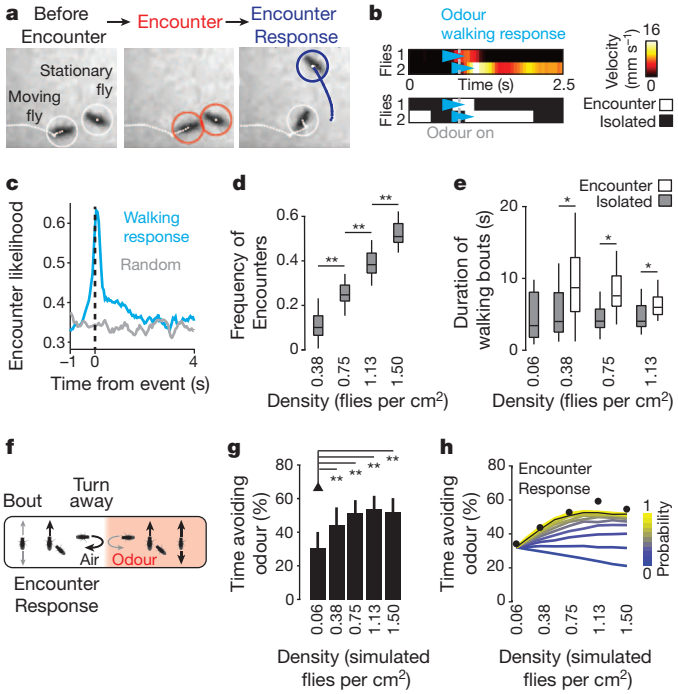


Figure 2 | Inter-fly Encounters coincide with odour responses and are required for collective odour avoidance in simulations. **a**, Images of two flies (left, white circles) undergoing an Encounter (middle, red circles) that results in an Encounter Response (right, blue circle). **b**, Velocities and Encounters for two flies exposed to CO_2 at odour onset (grey dashed line). The top panel shows the velocity for each fly. Cyan arrowheads indicate the first walking bout initiated after odour onset ('Odour walking response'). The bottom panel shows when these flies are (white) or are not (black) undergoing an Encounter during the same time period. **c**, The likelihood of an Encounter with respect to the time of the odour walking response (blue line) or a randomly chosen time point (grey line). Data are from Fig. 1d; density = 1.13 flies per cm^2 and $n = 200$ flies. **d**, The frequency of Encounters as a function of group density. Data are from Fig. 1d. **e**, The duration of walking bouts depending on whether they are initiated in isolation (grey boxes) or during an Encounter (white boxes). Data are from Fig. 1d. **f**, Simulated flies moved through a virtual arena as a function of three parameters: spontaneous bout probability ('Bout'), Encounter Response probability ('Encounter Response'), and turn away probability from the air–odour interface ('Turn away'). Low (small grey arrows) or high (large black arrows) probabilities were experimentally determined (Extended Data Fig. 2). **g**, The per cent time avoiding the odour (mean and s.d.) for five densities of simulated flies ($n = 80$ experiments for each condition). **h**, The sensitivity of simulated odour avoidance to Encounter Response probabilities ranging from 0 (never responding to Encounters, blue) to 1 (always responding, yellow). Each coloured line indicates the mean odour avoidance time ($n = 10{,}902$ experiments for each data point). The black line indicates Probability = 0.8, taken from real fly data in Fig. 1. Black circles indicate the mean fly avoidance times from Fig. 1d.

disc (Supplementary Video 4). We observed a stereotyped relationship between the location of mechanical touch and subsequent walking trajectories (Fig. 3d), whose associated kinematics were indistinguishable from those of Encounter Responses. Thus, mechanical touch alone can elicit Encounter Responses (compare Extended Data Fig. 3c and g; two-sample Kolmogorov–Smirnov test, $P = 0.3$). Consistently, genetic ablation of flies' oenocytes, to remove cuticular hydrocarbon contact chemosensory signals[17], had no effect on the ability of these animals to elicit Encounter Responses in wild-type flies (Fig. 3e). These data imply that Encounter Responses are mediated solely by mechanosensory stimulation.

We next identified mechanosensory neurons required for touch-evoked Encounter Responses by driving tetanus toxin (Tnt) expression with a panel of candidate mechanosensory Gal4 lines (Extended Data Fig. 4a). *R55B01-Gal4/UAS-Tnt* flies exhibited significantly diminished Encounter Responses compared to a gustatory neuron driver line (Extended Data Table 2), without reduced ability to produce sustained high-velocity walking bouts (Extended Data Fig. 4b). *R55B01-Gal4*-driven expression of a *UAS-CD4:tdGFP* reporter was detected in neurons innervating leg and wing neuropils of the thoracic ganglia (Extended Data Fig. 5a). Consistently, green fluorescent protein (GFP) labelled neurons in several leg mechanosensory structures: the femoral and tibial chordotonal organs, and distal leg mechanosensory sensilla neurons (Extended Data Fig. 5b). Notably, among the screened lines only *R55B01-Gal4* drove expression in leg mechanosensory sensilla (Extended Data Fig. 4c, d), suggesting that these are the critical neurons for Encounter Responses.

To ascertain the contribution to Encounter Responses of leg mechanosensory sensilla and/or chordotonal structures (which can also sense touch[18,19]), we identified additional Gal4 driver lines that drove expression in subsets of these neuron classes. By intersecting *piezo-Gal4* with *cha3-Gal80*, a Gal4 suppression line, we could limit leg expression to mechanosensory sensilla neurons (termed 'Mechanosensory Sensilla driver' line) (Fig. 3f). Importantly, silencing neurons with this driver significantly diminished Encounter Response frequency (Fig. 3g). By contrast, silencing leg chordotonal organs alone had no effect on Encounter Response frequency (Extended Data Fig. 5a–c).

We tested the sufficiency of leg mechanosensory sensilla neuron activity to elicit Encounter Response-like walking by expressing channelrhodopsin-2 (ChR2) in each class of leg mechanosensory neurons

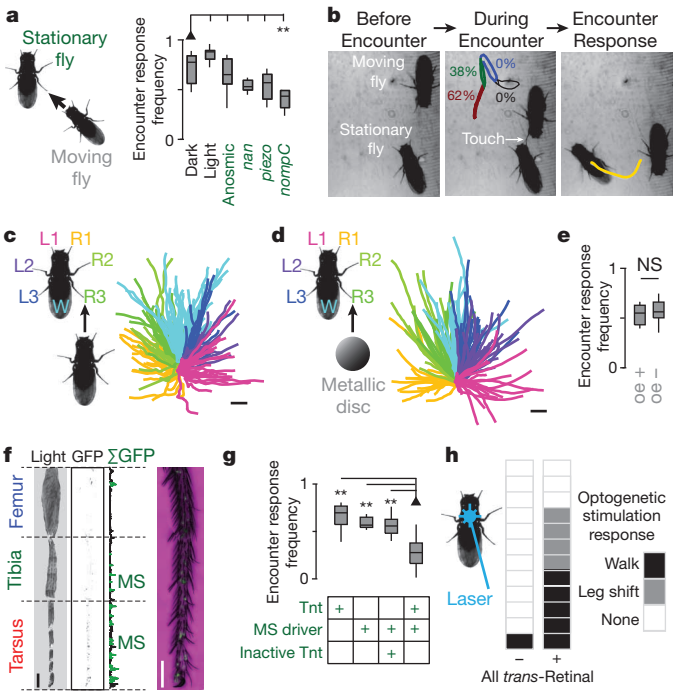


Figure 3 | Leg mechanosensory sensilla neuron activity is necessary and sufficient for Encounter Responses. **a**, The frequency of Encounter Responses measured from experiments in Fig. 1 ('Dark'), illuminated experiments ('Light'), near-anosmic mutants (*IR8a*^{1}, *IR25a*^{2}, *GR63a*^{1}, *ORCO*^{1}), auditory/proprioceptive mutants (*nanchung*^{36a}), nociceptive touch mutants (*piezo*^{KO}), and gentle touch mutants (*nompC*^{f00914}). To calculate the frequency of Encounter Responses, we tested how often each stationary fly undergoing an Encounter moved continuously for the next half-second. $n = 10$ experiments for each condition (density = 0.75 flies per cm^2). Reductions for *nanchung* and *piezo* mutants were not statistically significant (Extended Data Table 1). **b**, Single frames from a high-resolution video of an Encounter between a moving fly and a stationary fly. The schematic on the middle frame shows the per cent of all observed Encounter Responses resulting from touch for each leg segment ($n = 104$ experiments). The Encounter Response walking trajectory elicited by touch is shown in yellow on the right-hand frame. **c**, Encounter Response trajectories (right) colour-coded by the appendage touched by the neighbouring fly (left). Wings, W ($n = 54$ experiments); legs, R1–R3 and L1–L3 ($n = 21, 18, 19$ and $23, 15, 17$ experiments, respectively). The scale bar is 1 mm and each trajectory represents up to 0.24 s of walking. **d**, Touch response trajectories (right) colour-coded by which appendage was touched by a metallic disc (left). Wings, W ($n = 20$ experiments); legs, R1–R3 and L1–L3 ($n = 20, 21, 21$ and $18, 21, 20$ experiments, respectively). The scale bar is 2.5 mm and each trajectory represents up to 1.5 s of walking. **e**, The frequency of Encounter Responses elicited by moving flies with ('oe+') or without ('oe−') cuticular hydrocarbon-secreting oenocytes ($n = 11$ experiments each). NS, not significant. **f**, A transmitted light image, inverted fluorescence image (fluorescence in black), and summed fluorescence (ΣGFP) for a Mechanosensory Sensilla driver fly leg expressing GFP (*UAS-CD4:tdGFP/piezo-Gal4;cha3-Gal80/+*). Leg mechanosensory sensilla ('MS') are indicated in green. A high-resolution image of the tarsus is shown on the right. Endogenous GFP fluorescence (green) is superimposed upon a transmitted light image (magenta). The scale bars are 100 µm. **g**, The frequency of Encounter Responses for parental line controls (*UAS-Tnt/+;+* or *piezo-Gal4/+;cha3-Gal80/+*), Mechanosensory Sensilla driver flies expressing an inactive tetanus toxin control (*UAS-Tnt*^{IMP}*/piezo-Gal4;cha3-Gal80/+*), or Mechanosensory Sensilla driver flies expressing tetanus toxin (*UAS-Tnt/piezo-Gal4;cha3-Gal80/+*). $n = 12, 13, 15$ and 15 experiments, respectively. **h**, Blue laser optogenetic stimulation responses of flies expressing ChR2 in mechanosensory sensilla (*piezo-Gal4/+*; *cha3-Gal80/UAS-ChR2(T159C)*) in the absence (left) or presence (right) of the essential cofactor all *trans*-retinal ($n = 12$ flies for each condition). Each box indicates the response for a single fly ('walk', 'leg shift' or 'none').

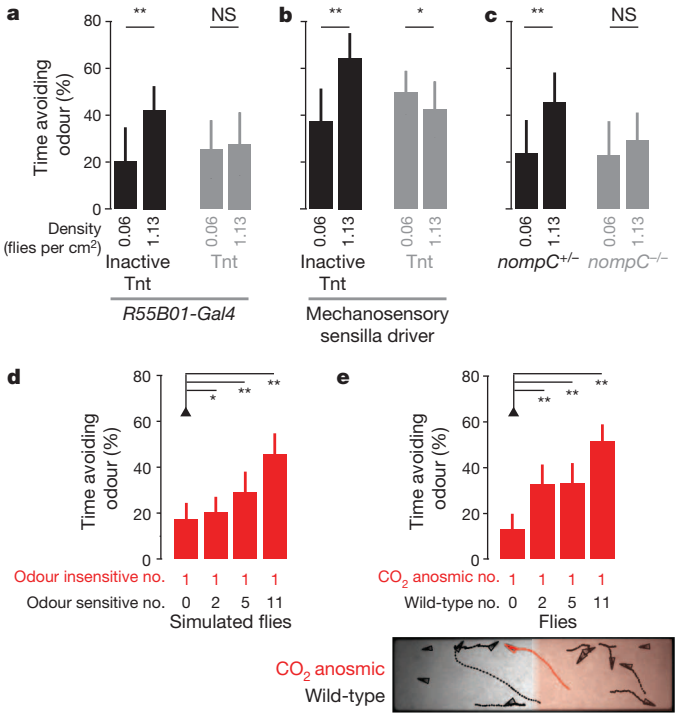


Figure 4 | Encounter Responses are necessary and sufficient for collective odour avoidance. **a**, **b**, The per cent time avoiding the odour (mean and s.d.) for *R55B01-Gal4* (**a**) or Mechanosensory Sensilla driver (**b**) flies expressing an inactive tetanus toxin control, or tetanus toxin. $n = 22, 21, 22$, and 19 experiments for *R55B01-Gal4* and $n = 23, 21, 21$, and 21 experiments for the Mechanosensory Sensilla driver (genotypes: *UAS-Tnt*^{IMP}*;R55B01-Gal4*, *UAS-Tnt;R55B01-Gal4*, *UAS-Tnt*^{IMP}*/piezo-Gal4;cha3-Gal80/+*, *UAS-Tnt/piezo-Gal4;cha3-Gal80/+*). **c**, The per cent time avoiding the odour (mean and s.d.) for heterozygous control, or homozygous *nompC*^{f00914} mutant animals. $n = 22, 22, 21$, and 21, respectively. **d**, **e**, The per cent time avoiding the odour (mean and s.d.) for individual CO_2 anosmic virtual and real flies (*GR63a*^{1}*,IR64a*^{MB05283}). Avoidance time is measured from a single 'CO_2 anosmic' fly per experiment in a simulated model (**d**, $n = 80$ experiments each), or in *Drosophila* (**e**, $n = 35, 37, 40$ and 38 experiments) where single mutant flies were tested for CO_2 avoidance in the context of wild-type flies.

and recording behavioural responses to blue light pulses. Optogenetic stimulation of flies expressing ChR2 in leg mechanosensory sensilla neurons, but not chordotonal organs, resulted in Encounter Response-like walking (Fig. 3h; Extended Data Fig. 5d, Supplementary Videos 5 and 6), consistent with natural elicitation of Encounter Responses by inter-fly touch of distal leg segments (Fig. 3b, inset).

Our identification of a neuronal basis for Encounter Responses allowed us to test our model's prediction (Fig. 2h) that inter-fly interactions are required for collective odour avoidance. First, we silenced leg mechanosensory sensilla neurons by expressing Tnt with *R55B01-Gal4* or the Mechanosensory Sensilla driver. Second, we studied *nompC* mutants. Each of these perturbations abolished collective odour avoidance (Fig. 4a–c), supporting the link between mechanosensation and group behaviour.

Touch may enhance odour avoidance by increasing awareness of the stimulus. Alternatively, touch may produce an odour-independent Encounter Response reaction that initiates departure from the odour zone. To distinguish between these possibilities, we asked if odour-insensitive flies displayed increased avoidance in the presence of odour-sensitive animals. Indeed, both in simulations (Fig. 4d) and in real flies (Fig. 4e), increasing the number of odour-sensitive individuals led to greater avoidance behaviour of odour-insensitive individuals. Thus, in this context, touch-mediated modulation of odour awareness plays little, if any, role in collective avoidance.

Combining systems-level and neurogenetic approaches, we have uncovered a hierarchy of mechanisms that drive collective motion in *Drosophila*. Active flies elicit spatially stereotyped walking responses in stationary flies through appendage touch interactions, requiring the NOMPC mechanosensory channel and distal leg mechanosensory sensilla neurons. Through Encounter Responses, odour reactions of sensitive flies spark cascades of directed locomotion of less sensitive (or even insensitive) individuals, causing a coherent departure from the odour zone. This behavioural positive feedback and group motion are absent among flies in the non-odour zone since they are less likely to initiate walking and, consequently, have a reduced frequency of Encounters. Additionally, flies retreat when encountering the odour while transiting from the air zone. Together these behavioural phenomena cause flies to escape the odour zone and then remain in the air zone, resulting in higher odour avoidance for groups compared to isolated animals (Extended Data Fig. 6). When distal appendage mechanosensory touch detection is impaired, groups of flies cannot produce Encounter Responses, are less likely leave the odour zone, and instead behave like isolated flies. Encounters are likely to have widespread influence on sensory-evoked actions of individuals in groups. For example, movement of flies towards areas of high elevation[20] is also increased in higher density groups (Extended Data Fig. 7).

Behaviour in animal groups arises from the detection and response to intentional and unintentional signals of conspecifics. While neural circuits controlling pairwise interactions, such as courtship, are increasingly well-understood[21], we know little about those orchestrating group-level behaviours. The identification of sensory pathways that mediate collective behaviour in *Drosophila* opens the possibility to understand the neural basis by which an individual's actions may influence—and be influenced by—group dynamics.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 15 July; accepted 29 October 2014.
Published online 24 December 2014.

1. Berdahl, A., Torney, C. J., Ioannou, C. C., Faria, J. J. & Couzin, I. D. Emergent sensing of complex environments by mobile animal groups. *Science* **339,** 574–576 (2013).
2. Ward, A. J., Herbert-Read, J. E., Sumpter, D. J. T. & Krause, J. Fast and accurate decisions through collective vigilance in fish shoals. *Proc. Natl Acad. Sci. USA* **108,** 2312–2315 (2011).
3. Couzin, I. D. Collective cognition in animal groups. *Trends Cogn. Sci.* **13,** 36–43 (2009).
4. Sumpter, D., Buhl, J., Biro, D. & Couzin, I. D. Information transfer in moving animal groups. *Theory Biosci.* **127,** 177–186 (2008).
5. Walker, R. G., Willingham, A. & Zuker, C. A *Drosophila* mechanosensory transduction channel. *Science* **287,** 2229–2234 (2000).
6. Yan, Z. *et al. Drosophila* NOMPC is a mechanotransduction channel subunit for gentle-touch sensation. *Nature* **493,** 221–225 (2013).
7. Gullan, P. J. & Cranston, P. S. *The Insects* (Wiley, 2010).
8. Schneider, J., Atallah, J. & Levine, J. D. One, two, and many—a perspective on what groups of *Drosophila melanogaster* can tell us about social dynamics. *Adv. Genet.* **77,** 59–78 (2012).
9. Levine, J. D., Funes, P., Dowse, H. B. & Hall, J. C. Resetting the circadian clock by social experience in *Drosophila melanogaster. Science* **298,** 2010–2012 (2002).
10. Wang, L. & Anderson, D. J. Identification of an aggression-promoting pheromone and its receptor neurons in *Drosophila. Nature* **463,** 227–231 (2010).
11. Suh, G. S. B. *et al.* A single population of olfactory sensory neurons mediates an innate avoidance behaviour in *Drosophila. Nature* **431,** 854–859 (2004).
12. Ai, M. *et al.* Acid sensing by the *Drosophila* olfactory system. *Nature* **468,** 691–695 (2010).
13. Billeter, J.-C. & Levine, J. D. Who is he and what is he to you? Recognition in *Drosophila melanogaster. Curr. Opin. Neurobiol.* **23,** 17–23 (2013).
14. Schneider, J., Dickinson, M. H. & Levine, J. D. Social structures depend on innate determinants and chemosensory processing in *Drosophila. Proc. Natl Acad. Sci. USA* **109,** 17174–17179 (2012).
15. Schaefer, P. L., Varuni Kondagunta, G. & Ritzmann, R. E. Motion analysis of escape movements evoked by tactile stimulation in the cockroach *Periplaneta americana. J. Exp. Biol.* **190,** 287–294 (1994).
16. Stocker, R. F. Taste perception: *Drosophila* – a model of good taste. *Curr. Biol.* **14,** R560–R561 (2004).
17. Billeter, J.-C., Atallah, J., Krupp, J. J., Millar, J. G. & Levine, J. D. Specialized cells tag sexual and species identity in *Drosophila melanogaster. Nature* **461,** 987–991 (2009).
18. Höltje, M. Rapid mechano-sensory pathways code leg impact and elicit very rapid reflexes in insects. *J. Exp. Biol.* **206,** 2715–2724 (2003).
19. Kamikouchi, A., Wiek, R., Effertz, T., Göpfert, M. C. & Fiala, A. Transcuticular optical imaging of stimulus-evoked neural activities in the *Drosophila* peripheral nervous system. *Nature Protocols* **5,** 1229–1235 (2010).
20. Hirsch, J. & Erlenmeyer-Kimling, L. Sign of taxis as a property of the genotype. *Science* **134,** 835–836 (1961).
21. Manoli, D. S., Fan, P., Fraser, E. J. & Shah, N. M. Neural control of sexually dimorphic behaviors. *Curr. Opin. Neurobiol.* **23,** 330–338 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank L. Sprecher, J. Weber, A. Gaille, A. Canapini and I. Barbier for help in aggregation density measurements, A. Silbering for generating anosmic *Drosophila* lines, F. Schütz for advice on statistics, J. Yi for image analysis software, J. Levine, A. Patapoutian, M. Landgraf, M. Göpfert and the Bloomington *Drosophila* Stock Center for *Drosophila* strains, T. Oertner for plasmids, and the Developmental Studies Hybridoma Bank for antibodies. We thank D. Cullen, L. Keller, J. Levine, M. Louis, S. Manley, S. Martin, J. Schneider and members of the Benton and Floreano laboratories for discussions. P.R. was supported by a Human Frontier Science Program Long-term Fellowship. P.L. was supported by the Swiss National Science Foundation (200021_127143). D.F. acknowledges support from the Swiss National Science Foundation (CR32I3_141063/1) and the FP7-FET European Project INSIGHT (308943). R.B. acknowledges support from European Research Council Starting Independent Researcher and Consolidator Grants (205202 and 615094).

Author Contributions P.R., R.B. and D.F. conceived, designed and supervised the project. P.R., S.C., P.L. and L.F. performed experiments. P.R., P.L., L.F. and W.T. analysed data. P.R. and R.B. wrote the paper with assistance from P.L. and D.F.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to R.B. (Richard.Benton@unil.ch) or P.R. (ramdya@gmail.com).

METHODS

***Drosophila* lines.** *Actin88F-eGFP* (ref. 22) backcrossed 5 generations to w^{1118} was used as the wild-type line enabling distinction from non-fluorescent mutant flies in Fluorescence Behavioural Imaging experiments (Fig. 3a,e,g and Fig. 4e and Extended Data Fig. 5c).

GR63a^{1},*IR64a*^{MB05283} mutant flies were used as the CO_2-anosmic individuals (Fig. 4e).

IR8a^{1};*IR25a*^{2};*GR63a*^{1},*ORCO*^{1} quadruple mutant flies were used to measure the influence of olfaction on the frequency of Encounter Responses (Fig. 3a).

nanchung^{36a} (ref. 23), *piezo*^{KO} (ref. 24) and *nompC*^{f00914} (ref. 5) mutant flies were used to measure the impact of mechanosensing on the frequency of Encounter Responses (Fig. 3a).

P{GMR-Gal4}attP2 transgenic flies[25,26] were used to identify neural populations with deficient touch responses (Fig. 3g, and Extended Data Figs 4,5). Gal4 drivers were selected by pre-screening a large panel (http://flweb.janelia.org/) for those displaying sparse expression in neurons that projected from the legs to the thoracic ganglia and neurons innervating the antennal mechanosensory and motor centre in the brain[26]. To identify *R55B01-Gal4*, we compared the frequency of Encounter Responses in animals bearing these transgenes against that of a control driver, *R27B07-Gal4*, which drives a gustatory pattern of expression in the legs[27] and Thoracic Ganglia (Extended Data Fig. 4a, green and data not shown). Brain expression in *R55B01-Gal4* was limited to neurons projecting to the Antennal Mechanosensory and Motor Centre, and weaker expression in those innervating the suboesophageal zone (which receives both gustatory and mechanosensory input from the labellum) and several visual areas (optic lobes and optic tubercle; Extended Data Fig. 5a). These weakly marked neural populations are likely to contribute only minimally to Encounter Responses as labellar touch was never observed and all experiments were performed in the dark. Finally, we also observed fan-shaped body expression in *R55B01-Gal4*. However, inhibiting the fan-shaped body cannot explain Encounter Response reductions in *R55B01-Gal4* since silencing these neurons alone has no effect on Encounter Response frequency (Extended Data Fig. 5c, *R65C03-Gal4*).

piezo-Gal4;cha3-Gal80 flies were used to target mechanosensory sensilla neurons.

UAS-Tnt and *UAS-Tnt*^{IMP} flies were used to measure the effects of neural knockdown on Encounter Responses and collective behaviour (Figs 3g, 4a, b, and Extended Data Fig. 5c).

UAST-ChR2(T159C) flies were generated by cloning ChR2(T159C) (ref. 28) into pgUASTgattB (refs 29, 30) and inserting this transgene into attP2 site (Genetic Services, Inc., Cambridge MA, USA). *UAS-ChR2* flies were then crossed with Gal4 driver lines for channelrhodopsin-2 stimulation experiments (Fig. 3h, and Extended Data Fig. 5d).

PBac{y[+mDint2] w[+mC] = UAS-CD4:tdGFP}VK00033 flies[31] were used to visualise Gal4 driver expression in leg, brain, and thoracic ganglia neurons (Fig. 3f and Extended Data Figs 4d, e and 5a, b).

PromE(800)-Gal4 [4M],Tub:Gal80ts flies, *UAS-StingerII, UAS-Hid/CyO* flies and *UAS-StingerII* flies were used for oenocyte ablation experiments (Fig. 3e) as described previously[32].

Experimental and statistical conditions. Experiment sample sizes were chosen based on preliminary studies. If sample size constraints and proper experimental conditions were met, all experiments were included for subsequent analysis. Experiments for different conditions and genotypes were interleaved to minimise the effects of time-of-day on behavioural results. Owing to the automated nature of almost all data acquisition and analysis, the experimenter was not blinded. For data meeting the criteria of normality, bar plots are presented and parametric statistical tests were used. For other data, boxplots and non-parametric statistics were used. Groups with similar variance are compared throughout the study.

Arena design and flow simulation. Arenas were designed using the 3D CAD software, SolidWorks (Dassault Systemes, Waltham, Massachusetts, USA) and CNC machined from polyoxymethylene and acrylic glass. Arena flow patterns were simulated using EasyCFD (http://www.easycfd.net) incorporating measured physical and flow parameter values (Extended Data Fig. 1d).

Behavioural imaging and tracking. For low-resolution behavioural imaging, we used Fluorescence Behavioural Imaging (FBI)[22] acquisition software and hardware. In all cases, we used Ctrax[33] for fly tracking and data analysis was performed using custom Matlab scripts (The Mathworks, Natick, Massachusetts, USA).

Behavioural experiments. All experiments were performed on adult female *Drosophila* raised at 25 °C on a 12 h light:12 h dark cycle 2–4 days post-eclosion, with the exception of experiments in Extended Data Fig. 1b, which used male flies). Experiments were performed in a temperature-controlled room at 25 °C, except for those in Fig. 3d,h, which were performed at 22 °C. In all cases except for aggregation measurements, flies were starved in empty 50 mm Petri dishes for 3–6 h in humidified 25 °C incubators. Experiments were performed in either the morning or late afternoon, Zeitgeber time.

Aggregation measurements (Extended Data Fig. 1). Flies were starved for 24 h at 25 °C in tubes humidified with moist Kimwipes. Ripe banana paste was prepared on the day of experiments and placed into a 12.78 cm^2 dish. Experiments were performed in either the morning or late afternoon Zeitgeber time. Experiments for summary data (Extended Data Fig. 1b) were filmed with a webcam (Microsoft LifeCam Studio, Redmond, USA) for 90 min with images acquired every 10 min. Flies were placed into a clean transparent box (23.5 cm × 25 cm × 37.5 cm) with only red light illumination. To calculate densities, the number of flies on the food source was calculated for each image and averaged from the 30th to 90th minute.

Collective odour avoidance – wild-type, neural knockdown, *nompC*, and mixed wild-type/anosmic (Fig. 1 and Fig. 4a–c,e). For olfactory stimulation, premixed 5% CO_2, or air (Messer Schweiz AG, Lenzburg, Switzerland) was flowed through Mass Flow controllers (PKM SA, Lyss, Switzerland) at a regulated flow rate of 500 ml min^{-1} via computer controlled solenoid valves (The Lee Company, Westbrook, CT, USA). A custom-fabricated circuit board and software[22] (sQuid, http://lis.epfl.ch/squid/) controlled valves, illumination LEDs (Super Bright LEDs Inc. St Louis Missouri, USA), and acquisition cameras (Allied Vision Technologies, Stadtroda, Germany). Flies were imaged in the olfactory arena using the following illumination/olfactory stimulation protocol: (1) infrared/blue light; air both sides (10 s); (2) infrared light, 5% CO_2/air (2 min); (3) infrared/blue light; air both sides (10 s).

The arena half with CO_2 was varied across experiments to eliminate the effects of other possible environmental asymmetries on behavioural results. Blue light was used in all cases to keep experiments consistent with mixed genotype FBI collective behaviour experiments (Fig. 4e).

Collective negative gravitaxis (Extended Data Fig. 7). For negative gravitaxis experiments, we tilted the behavioural arena at a 22.5° incline for 2 min. Flies were placed near the lower portion of the arena and were illuminated with red light. The Negative Gravitaxis Index was calculated by averaging their position along the long axis of the arena (with values ranging from 0 (bottom of the arena) to 100 (highest point of the arena)) during the second minute of the experiment.

Encounter Response modality screen (Fig. 3a). 12 wild-type flies ('light') or mixtures of 6 wild-type and 6 mutant flies (using a GFP reporter and Fluorescence Behavioral Imaging to distinguish genotypes[22]) were imaged in the olfactory arena using the following illumination/odorant protocol: (1) infrared/blue light; air both sides (10 s); (2) infrared light, air both sides (5 min); (3) infrared/blue light, air both sides (10 s).

Blue light was used in all cases to keep experiments consistent with mixed genotype Encounter Response experiments.

High-resolution inter-fly touch response (Fig. 3b, c and Extended Data Fig. 3d, e). Four flies were imaged in a small arena (1 cm × 5 cm) backlit with infrared light (Super Bright LEDs Inc. St Louis Missouri, USA). Images were continuously acquired at 125 frames per second (fps) using a high-speed video camera (Fastec Imaging, San Diego, CA, USA). The experimenter captured a video if a stationary fly exhibited touch-elicited walking.

High-resolution mechanical touch response (Fig. 3d and Extended Data Fig. 3f, g). Individual flies were imaged in a small arena (3 cm × 3 cm) illuminated by a red ring light (FALCON Illumination MV, Offenau, Germany). Images were continuously buffered at 20 fps using a high-resolution video camera (Gloor Instruments, Uster Switzerland). A small magnetic metallic disc (1 mm diameter) was directed to individual leg or wing appendages using a larger permanent magnet. The experimenter captured a video if a stationary fly exhibited touch-elicited walking.

Neural silencing Encounter Response screen (Extended Data Fig. 4a, b). 18 flies expressing inactive Tnt, or Tnt under the control of a specific Gal4 driver were imaged in the group arena using the following illumination protocol: (1) infrared/blue light (10 s); (2) infrared light (2 min); (3) infrared/blue light (30 s); (4) infrared light (2 min); (5) infrared/blue light (10 s).

Neural silencing Encounter Response frequency (Fig. 3g and Extended Data Fig. 5c). 6 flies expressing *UAS-Tnt/+*,*<driver>-Gal4/+*, *UAS-Tnt/<driver>-Gal4*, or *UAS-Tnt*^{IMP}*/<driver>-Gal4* were imaged in the presence of 6 wild-type flies in the group arena using the following illumination protocol: (1) infrared/blue light; air both sides (10 s); (2) infrared light, air both sides (2 min); (3) infrared/blue light, air both sides (10 s).

Optogenetic stimulation (Fig. 3h and Extended Data Fig. 5d). Flies bearing *UAS-ChR2* (*T159C*) and the specified Gal4 driver were raised either in food mixed with 2 mM all *trans*-Retinal ('ATR', Sigma-Aldrich, St Louis USA) or in the 95% ethanol solvent. Individual flies (2-4 days post-eclosion) were imaged in a small arena (3 cm × 3 cm) illuminated by a red ring light (FALCON Illumination MV, Offenau, Germany). Images were continuously buffered at 20 fps using a high-resolution video camera (Gloor Instruments, Uster Switzerland). An optically coupled red laser (Thorlabs, Newton, USA) was aligned to target the fly's thoracic segment. Stimulation consisted of a short (1 s) pulse of blue laser light (Coherent,

Santa Clara, USA). The experimenter video recorded up to three stimulations per fly at a spacing of approximately 2 min; scored responses were observed at least twice.

Behavioural analysis. To determine threshold values for fly motion, Encounters, and Encounter Responses, we measured velocities, accelerations and distances that could conservatively account for a test data set of manually annotated events. To the best of our knowledge our results are qualitatively robust to small variations in these values.

Percent of time avoiding odour (Fig. 1d, e, Fig. 2g, h, Fig. 4, and Extended Data Fig. 2e–h). To calculate odour avoidance, we measured the per cent of time that flies spent in the non-odour (air) zone during the experiment's second minute. This time period was chosen since we observed that flies tend to reduce exploration after one minute; see Fig. 1f. '% of time avoiding odour' = (time spent in the odour zone during the last minute/1 min) × 100. To report quantitatively equivalent values for experiments with different densities of flies, we resampled data using bootstrapping. This entailed randomly selecting a subgroup of experiments and, from these, one fly per experiment. We then averaged the odour avoidance for these flies to yield one result. We repeated this process a specified number of iterations to generate a distribution from which to report the mean and s.d. The number of iterations was closely linked to the average number of experiments. For example, in cases where $n \approx 40$, the number of bootstrapping iterations was 40.

Walking bouts (Fig. 2e and Extended Data Fig. 2b). We measured activity bouts using a hysteresis threshold on forward velocity (Extended Data Fig. 2b) or velocity magnitude (Fig. 2e) to create a binary time-series. Bouts began when velocity exceeded a high threshold of 1 mm s^{-1}. Bouts ended when velocity was below a low threshold of 0.5 mm s^{-1}. Short bouts or pauses (<2 frames or 100 ms, see Extended Data Fig. 2b; <20 frames or 1 s, see Fig. 2e) were removed by merging the fly's state with neighbouring measurements. Bouts were also terminated when moving flies encountered obstacles including other flies. This can explain the decreasing Encounter induced bout lengths observed at higher densities (Fig. 2e).

Coherent motion index (Extended Data Fig. 1g, h). To measure the coherence of group motion away or towards the odour zone, we calculated a coherent motion index (CMI). We did this by first identifying walking flies at every time-point. For these flies, we identified the orientation of walking in a binary fashion: within the half-circle pointing towards the odour half of the arena or within the half-circle pointing towards the air half of the arena. The CMI for each time-point is: (no. of flies moving towards the air – no. of flies moving towards the odour)/total no. of moving flies.

For a given experimental replicate, we average the CMI for the first ten seconds of odour presentation to capture the initial avoidance response. We report the distribution of this time-averaged CMI value across experimental replicates. For our analysis we examined the CMI for flies starting either in the air zone or the odour zone. Since the number of flies in a replicate can affect possible CMI values, comparisons should be limited to experiments with the same density of flies.

Encounter likelihood/frequency of Encounters (Fig. 2b–d). To calculate Encounter likelihood with respect to odour walking responses (Fig. 2b, c), we identified odour reactions as the time at which a stationary fly within the odour zone began moving (velocity magnitude >1 mm s^{-1}). As a control, a random time was selected from the entire experiment. We then determined the times at which each fly was undergoing an Encounter (distance to nearest neighbour <25% long-axis body length). Using these two data sets, we performed an event-triggered average of the Encounter time-series for all flies.

Notably, the timing of the peak in Encounter likelihood is not of sufficient resolution to make inferences about causality. This is due to the inability to precisely define a touch encounter in low-resolution video for which the legs are not visible. With Encounters, we instead rely on an estimate based on the overlap between two circles defining the peripheral space of neighbouring flies. Therefore Encounters can continue past the onset of motion since neighbouring flies may not have become distant enough to terminate the Encounter. This is illustrated in Fig. 2b in which the Encounters (white blocks) persist past the times of 'odour walking response' (blue arrowheads) for both flies shown.

To calculate the frequency of Encounters for different group densities (Fig. 2d), we measured the proportion of time flies spent having Encounters during a given experiment. Notably, Encounters are a function of motion: flies that move are more likely to Encounter other flies.

Encounter Response frequency (Fig. 3a,e,g, Extended Data Fig. 4a, and Extended Data Fig. 5c). To calculate the Frequency of Encounter Responses, for each stationary fly (velocity magnitude <1 mm s^{-1}) undergoing an Encounter (distance to nearest neighbour <25% long-axis body length), we identified motion events (velocity >1 mm s^{-1} or angular velocity >2 rad s^{-1} or acceleration magnitude >15 mm s^{-2}). If there was continuous motion for the next half-second (mean velocity magnitude >5 mm s^{-1}) an Encounter Response occurred, otherwise not. The average frequency across all flies in a given experiment was used to calculate summary data. Notably, the Encounter Response frequency is normalized by the number of Encounters: Frequency of Encounter Responses = Encounters producing walking reaction/(Encounters producing walking reaction + Encounters eliciting no reaction). Therefore this frequency is not a function of motion. For example, flies with high walking probabilities may generate more Encounters but reactions to these interactions—Encounter Responses—may still be more or less frequent. Similarly, flies that are predominantly stationary may have few Encounters but these too may result in a high or low frequency of Encounter Responses.

Encounter Response trajectories and kinematics (Extended Data Fig. 3b, c). To calculate Encounter Response trajectories (Extended Data Fig. 3b), for each stationary fly (velocity magnitude <1 mm s^{-1} and angular velocity <2 rad s^{-1}) undergoing an Encounter (distance to nearest neighbour <25% long-axis body length) near the centre of the arena (distance to wall >2 mm), we identified motion events (angular velocity ≥2 mm s^{-1} or acceleration magnitude ≥15 mm s^{-2}). The position of the fly was recorded for the remaining frames until it stopped (velocity <1 mm s^{-1}) or became close to a new fly (distance to nearest neighbour <25% long-axis body length) or to a wall (distance to wall <2 mm). Resulting response trajectories were pooled across experiments as a function of the octant of the Encounter (Extended Data Fig. 3a; the appropriate octant was identified as the region surrounding the fly that was bisected by a straight line between the fly's centre of mass and that of the neighbouring fly). Encounter Response velocities were obtained for each of these trajectories and averaged to produce kinematic data. Boxplots were calculated by averaging over the first 500 ms of kinematic data (Extended Data Fig. 3c).

Touch response trajectories and kinematics (Fig. 3c, d, Extended Data Figs 3d–g). Trajectories were taken from raw tracking data of flies responding to touch. Trajectories ended when flies were near another fly or a wall. Each resulting response trajectory was pooled across experiments depending on the location of touch (for example, leg or wing). Touch response velocities were also obtained for each of these responses and averaged to recover kinematics. Boxplots were calculated by averaging over the first 160 ms (Extended Data Fig. 3e) or 500 ms (Extended Data Fig. 3g) of kinematic data. This discrepancy is due to the difference in frame-rate between the two measurements.

Comparing response kinematics. Our aim was to compare the shape of kinematic data across Encounter responses, interfly touch responses, and mechanical touch responses. However, these data could be quite distinct with regards to spatial and temporal resolution. Therefore, we first concatenated the median value from each of the common seven octants (excluding the front octant in the Encounter response data set) across each of three velocity measures (forwards, sideways, and angular velocities) yielding a vector with 21 data-points. We then normalized these 21 element vectors to range from 0 to 1. These vectors were then compared using the 2-sample Kolmogorov–Smirnov test.

Simulations

Simulated flies. To verify our model of collective odour avoidance we used an agent-based simulation driven by probabilistic behaviours (Fig. 2f–h, Extended Data Fig. 2). The artificial flies had a circular body of 2.5 mm diameter and were placed in the arena of size 80 mm × 20 mm for 2,400 time-steps (corresponding to 120 s of 'simulated' time). The odour was presented on one half of the arena during the entire simulation. Simulated flies walked with a constant speed of 0.51 mm per time-step in straight bouts, which were separated by periods of inactivity. At the beginning of each bout or when encountering an obstacle (a wall or another fly) each fly randomly changed its walking direction. The bouts were initiated either spontaneously in isolation or during an Encounter.

Isolated bouts (Extended Data Fig. 2a, b). To estimate the propensities of flies to initiate walking in isolation, we performed 45 additional single fly experiments (density = 0.06) in which individual animals walked in the dark for 2 min. Flies were exposed to air throughout the entire arena in the first minute and odour during the second minute. For each fly i (i = 1,2, ... 45), we integrated the differences between its consecutive positions during the first minute of the experiment (air) and separately during the second minute of the experiment (odour) at 20 Hz. The minimum of these 45 × 2 values (that is, 29.9 mm) was treated as accumulated noise and subtracted from all 90 values. Consequently, we obtained the total distance travelled in air and in odour by each of the 45 flies (that is, $D_{\mathrm{Air}}{}^{i}$ and $D_{\mathrm{Odour}}{}^{i}$ for i = 1, 2, … 45). To estimate bout durations, we rescaled these 90 values such that their mean was equal to the mean duration of Isolated bouts observed in the 'six flies' experiments (density = 0.38). Overall, we obtained 45 values of prototypical Isolated bout lengths initiated spontaneously in air and 45 values of prototypical Isolated bout lengths initiated in odour (that is, $L_{\mathrm{Air}}{}^{i} = 0.29D_{\mathrm{Air}}{}^{i}$ and $L_{\mathrm{Odour}}{}^{i} = 0.29D_{\mathrm{Odour}}{}^{i}$ for i = 1, 2, … 45). Of note, the estimated bout lengths varied between animals, and between air and odour for a single animal.

We used the 45 values of $L_{\mathrm{Air}}{}^{i}$ and the 45 values of $L_{\mathrm{Odour}}{}^{i}$ to bootstrap the behaviour of simulated flies. A simulated fly performed a self-induced bout of length $L_{\mathrm{Air}}{}^{s}$ if initiated in air, and of length $L_{\mathrm{Odour}}{}^{s}$ if initiated in odour, where s denoted which prototypical behaviour the simulated fly used. The value of s was set at time-step 1 for each simulated fly independently and uniformly at random to an

integer value between 1 and 45. Thus, the values of s varied between the flies but it was possible for some flies to have the same s value. The values of s were kept constant for each simulated fly during all 2,400 time-steps. Consequently, each simulated fly had a fixed propensity to move spontaneously. Moreover, within the same group, simulated flies usually differed in their propensity to move spontaneously.

Between bouts, a simulated fly following the prototypical behaviour s remained inactive for $(L_{Max} - {L_{Air}}^s)/v$ time-steps when resting in air and for $(L_{Max} - {L_{Odour}}^s)/v$ time-steps when resting in odour, where L_{Max} is the maximal value of all 90 values of L (that is, $L_{Max} = 214$ mm) and v is the walking speed. We estimated the walking speed v as the maximum of $({D_{Air}}^i + {D_{Odour}}^i)$ over all $i = 1,2, \ldots 45$ divided by 120 s, which resulted in $v = 10.2\ \text{mm s}^{-1} = 0.51$ mm per time-step.

Crossing the air–odour interface (Extended Data Fig. 2c). A simulated fly changed its direction of motion by 180 degrees when crossing from air to odour with probability P(turn away from odour) = 0.4, and when crossing from odour to air with probability P(turn away from air) = 0.2. The values of P(turn away from odour) and P(turn away from air) were estimated from 40 single fly (density = 0.06) experiments taken from Fig. 1d in which animals walked freely in the dark for 2 min with odour exposure on one half of the arena. We calculated the time flies spent in the odour after crossing from air, and vice versa. We classified a crossing from one half of the arena to another as a 'turn around' if the time spent in the new half was ≤ 3 s. Overall, we observed 76 crossings from air to odour out of which 31 were classified as a 'turn around', and 72 crossings from odour to air out of which 16 were classified as a 'turn around'.

Encounter-induced bouts (Extended Data Fig. 2d). In the simulation, at each time step t and for each walking fly we detected if the fly encountered an obstacle. If so, we checked whether in time-step $t + 1$ the fly's body would overlap with a wall or with other flies' bodies (assuming the fly would walk for 0.51 mm in the same direction it was heading). In these cases, the walking fly did not move in time-step t, but randomly changed its direction, and resumed the walk in time-step $t + 1$.

Moreover, if the walking fly encountered an inactive fly, it caused the encountered fly to initiate a bout with probability P(Encounter Response) = 0.8 (from Fig. 3a). The length of this Encounter Response bout was equal to $E({L_{Air}}^u)$ and to $E({L_{Odour}}^u)$ when initiated in air and in odour, respectively. The value u is a random integer between 1 and 45, and E is a mapping from the lengths of Isolated bouts to the lengths of Encounter Response bouts. Note that in contrast to the lengths of Isolated bouts (that is, ${L_{Air}}^s$ and ${L_{Air}}^s$) where the value s was fixed for each fly at the beginning of a simulation, here u was a random variable redrawn independently for each Encounter Response. Consequently, simulated flies did not vary in their propensity to move due to Encounters.

We did not explicitly encode directionality in the Encounter Response angle. However, we observed that since virtual flies cannot occupy the same space, stationary flies would move on average away from the location of touch, an implicitly directional response.

We estimated E using the data from the six fly experiments (density = 0.38) in which animals walked freely in green light illumination for 5 min without odour exposure. We observed 1,314 Isolated bouts and 618 Encounter-induced bouts. For all 1,932 bouts we calculated their lengths by integrating with temporal resolution of 20 Hz the differences between the consecutive positions of a given fly. Next, for both types of bouts, we calculated the 0th, 1st, 2nd, ... 100th percentiles of their lengths, created a scatter plot and calculated a double linear mapping from the lengths of Isolated bouts to the lengths of Encounter Response bouts, which fit the data best (that is, $E(x) = 4.71x + 0.75$ when $x < 20$ and $E(x) = 1.04x + 70.69$ otherwise).

Experiments and sensitivity analyses (Fig. 2h and Extended Data Fig. 2e–h). Overall, there were six experiments. We performed one main experiment to test the collective behaviour of flies in two conditions: (1) all flies in the group were odour-sensitive (Fig. 2h) and (2) the first fly from the group was odour-insensitive (Fig. 4d). To simulate an odour-insensitive fly s we used ${L'_{Odour}}^s = {L_{Air}}^s$ in place of ${L_{Odour}}^s$ values. Additionally, we performed five experiments, each corresponding to a different sensitivity analysis. For each of these experiments there were eleven conditions, each corresponding to a different value of the investigated parameter.

In the first experiment we varied the propensity to move due to Encounters by setting P(Encounter Response) between 0 and 1 with a step size of 0.1 (Fig. 2h). In the second experiment we varied the propensity to move in air (Extended Data Fig. 2e). To this end, we used ${L'_{Air}}^s = a_{Air}{L_{Air}}^s$ in place of ${L_{Air}}^s$ value, and we set the damping coefficient a_{Air} from 0 to 1 with a step size of 0.1. In the third experiment we varied the propensity to move in odour (Extended Data Fig. 2f). To this end, we used ${L'_{Odour}}^s = a_{Odour}{L_{Odour}}^s$ in place of ${L_{Odour}}^s$ value, and we set the damping coefficient a_{Odour} from 0 to 1 with a step size of 0.1. In the fourth experiment we varied the probability to turn back when crossing the interface from odour to air by setting P(turn away from air) between 0 and 1 with a step size of 0.1 (Extended Data Fig. 2g). In the fifth experiment we varied the probability to turn back when crossing the interface from air to odour by setting P(turn away from odour) between 0 and 1 with a step size of 0.1 (Extended Data Fig. 2h).

Odour avoidance. In both conditions of the main experiment, and for each pair of 5 sensitivity experiments and 11 conditions, we ran simulations with 9 different group sizes. We used groups composed of $n = 1$, 3, 6, 9, 12, 15, 18, 21, and 24 simulated flies (not all data reported). Overall, there were $1 \times 2 \times 9$ (main) and $5 \times 11 \times 9$ (sensitivity) lines of experiments. Each experimental line was replicated 22,000 times using a Mersenne Twister pseudo-random numbers generator[34] with a seed set to 1, 2, … 22,000, respectively. The initial positions, initial directions and the prototypical behaviours of simulated flies were identical between corresponding replicates across different experimental lines.

The odour avoidance of a simulated fly was calculated as the proportion of time-steps the fly spent in air during time-steps 1,200 to 2,400 corresponding to the second minute of the experiment. To compare simulations' outcomes between treatments, conditions and group sizes, we averaged in each experimental line the odour avoidance of the first simulated fly across all replicates in which the fly was initially placed in odour (there were 10,902 such replicates out of all 22,000 replicates). Note that we chose to compare experimental lines based on the first simulated fly because it was the only fly used in all experimental lines. For example, the second and the third simulated flies were present in all experimental lines with groups composed of 3 or more flies, but were not present when the group was composed of just one fly.

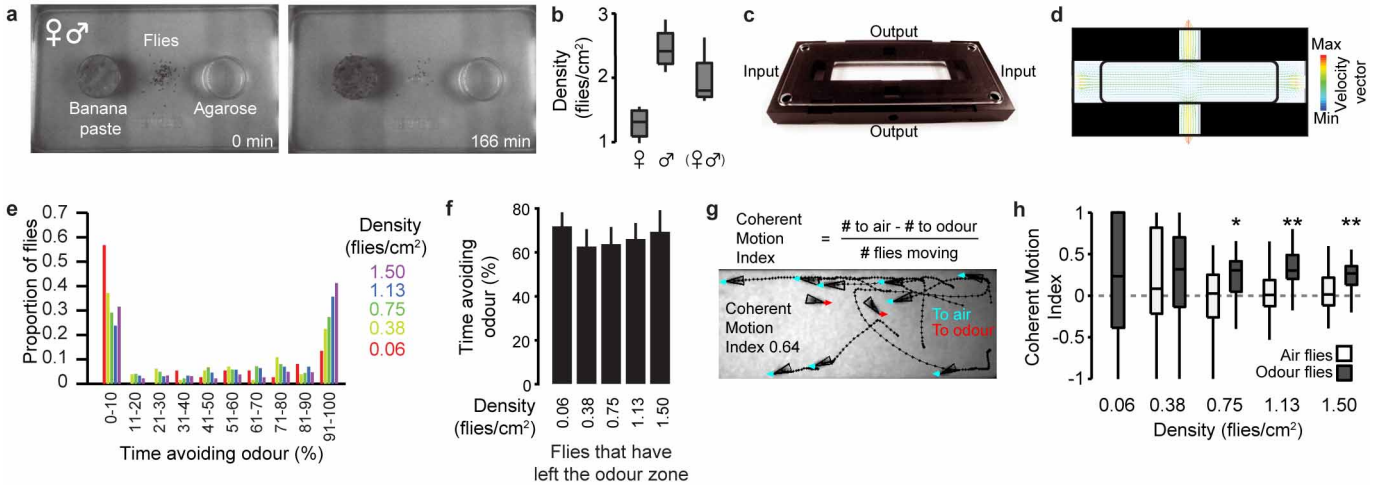
For more details see the simulation's implementation in Java available on-line at: http://documents.epfl.ch/users/r/ra/ramdya/www/ramdya/collective_sim.html.

Anatomical imaging

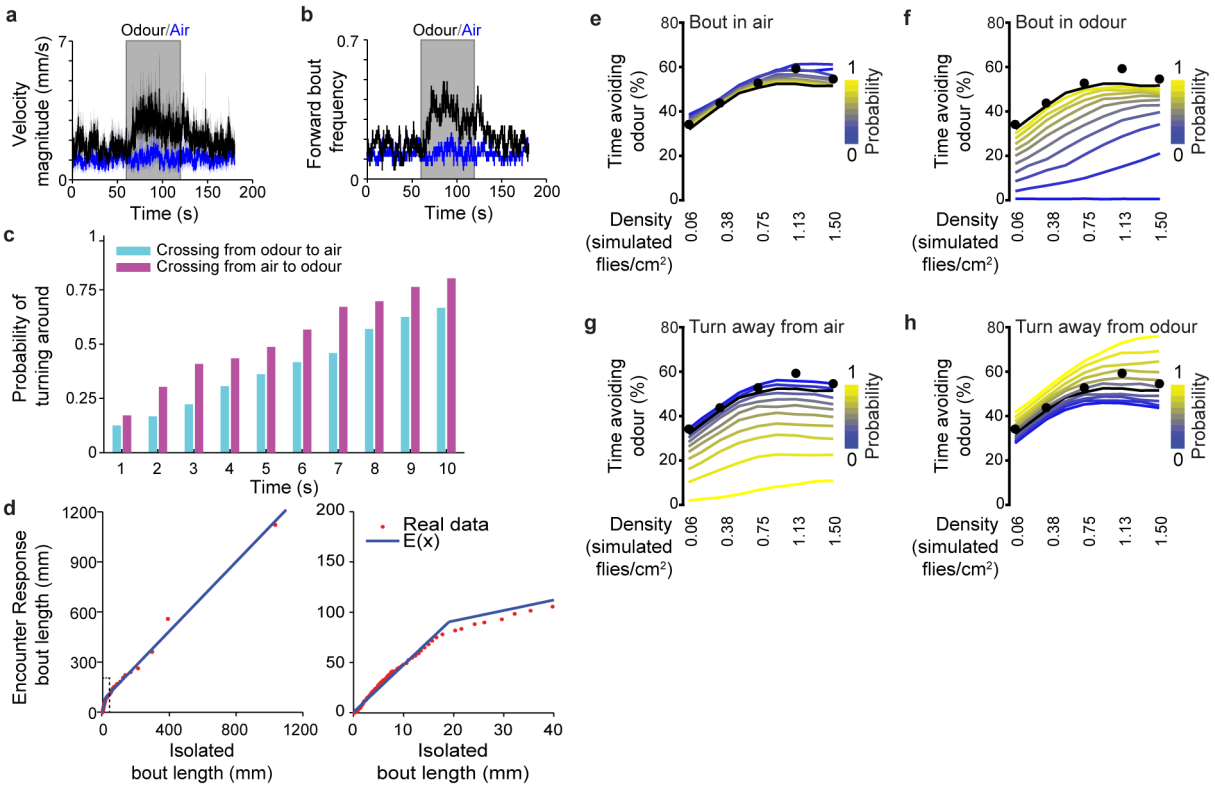
Brain/thoracic ganglia staining and imaging (Extended Data Fig. 5a). Immunofluorescence on whole-mount brains and thoracic ganglia was performed as described previously[29]. The primary antibodies were mouse monoclonal nc82 (1:10 dilution; Developmental Studies Hybridoma Bank), rabbit anti-GFP (1:200, Invitrogen A-6455). The secondary antibodies were Alexa Fluor 488- and Cy3- conjugated goat anti-rabbit or anti-mouse IgG, respectively (Molecular Probes and Jackson ImmunoResearch) diluted 1:250. Microscopy was performed using an LSM 510 laser scanning confocal microscope (Zeiss).

Leg neuron imaging (Fig. 3f and Extended Data Figs 4c, d and 5b). Legs were removed from female adults 2 days post-eclosion and mounted in VectaShield under a coverslip. Cuticle was imaged with a 543 nm laser while CD4:tdGFP was imaged using a 488 nm laser. Microscopy was performed using an LSM 510 laser scanning confocal microscope (Zeiss). We reoriented leg images using a custom script to identify and crop the femur, tibia, and tarsal segments. Using these sub-images, we then quantified fluorescence values (excluding autofluorescence from the cuticle and surface debris) orthogonal to the long axis of each leg segment to produce a profile of leg mechanosensory structures. Chordotonal organs and mechanosensory sensilla neurons were distinguished by morphology: sensilla neurons had small somata with dendrites projecting to the base of leg sensilla (Extended Data Fig. 4d).

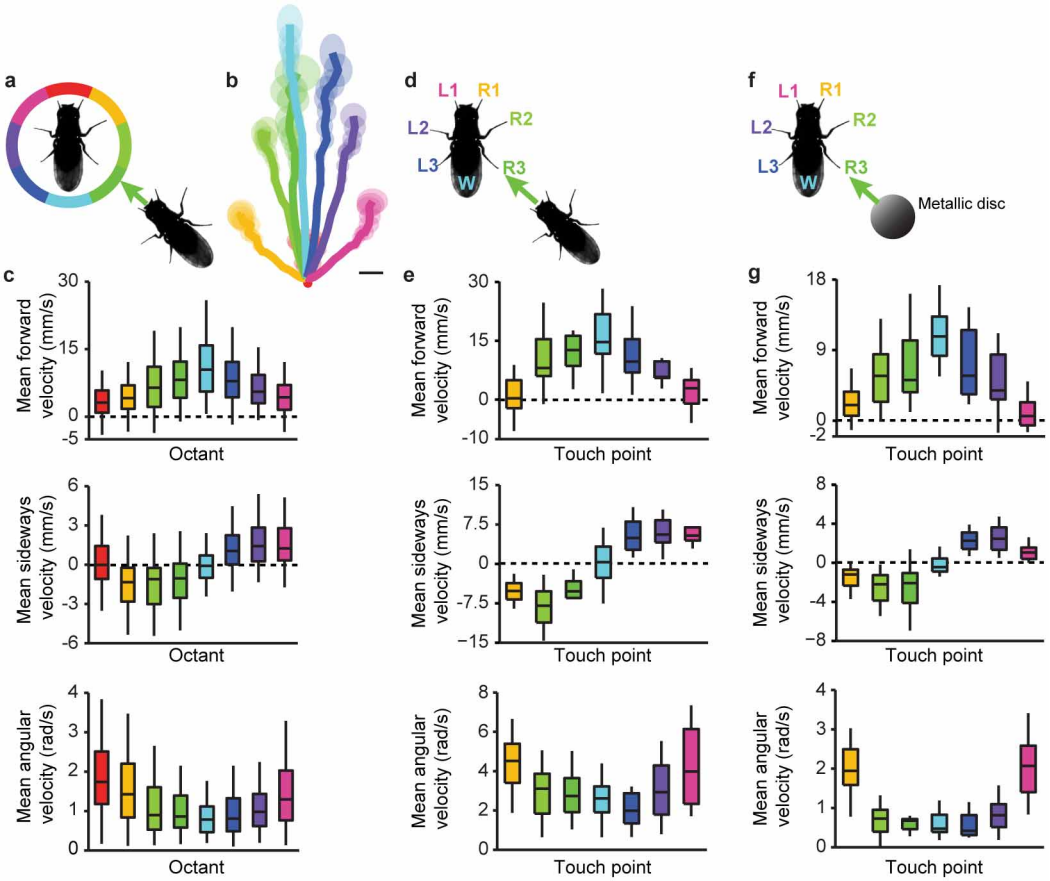
22. Ramdya, P., Schaffter, T., Floreano, D. & Benton, R. Fluorescence behavioral imaging (FBI) tracks identity in heterogeneous groups of *Drosophila*. *PLoS ONE* **7,** e48381 (2012).
23. Kim, J. *et al.* A TRPV family ion channel required for hearing in *Drosophila*. *Nature* **424,** 81–84 (2003).
24. Kim, S. E., Coste, B., Chadha, A., Cook, B. & Patapoutian, A. The role of *Drosophila* Piezo in mechanical nociception. *Nature* **483,** 209–212 (2012).
25. Brand, A. H. & Perrimon, N. Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. *Development* **118,** 401–415 (1993).
26. Jenett, A. *et al.* A GAL4-driver line resource for *Drosophila* neurobiology. *Cell Rep.* **2,** 1–11 (2012).
27. Ling, F., Dahanukar, A., Weiss, L. A., Kwon, J. Y. & Carlson, J. R. The molecular and cellular basis of taste coding in the legs of *Drosophila*. *J. Neurosci.* **34,** 7148–7164 (2014).
28. Berndt, A. *et al.* High-efficiency channelrhodopsins for fast neuronal stimulation at low light levels. *Proc. Natl Acad. Sci. USA* **108,** 7595–7600 (2011).
29. Silbering, A. F. *et al.* Complementary function and integrated wiring of the evolutionarily distinct *Drosophila* olfactory subsystems. *J. Neurosci.* **31,** 13357–13375 (2011).
30. Bischof, J., Maeda, R. K., Hediger, M., Karch, F. & Basler, K. An optimized transgenesis system for *Drosophila* using germ-line-specific C31 integrases. *Proc. Natl Acad. Sci. USA* **104,** 3312–3317 (2007).
31. Han, C., Jan, L. Y. & Jan, Y. N. Enhancer-driven membrane markers for analysis of nonautonomous mechanisms reveal neuron–glia interactions in *Drosophila*. *Proc. Natl Acad. Sci. USA* **108,** 9673–9678 (2011).
32. Billeter, J.-C., Atallah, J., Krupp, J. J., Millar, J. G. & Levine, J. D. Specialized cells tag sexual and species identity in *Drosophila melanogaster*. *Nature* **461,** 987–991 (2009).
33. Branson, K., Robie, A. A., Bender, J., Perona, P. & Dickinson, M. H. High-throughput ethomics in large groups of *Drosophila*. *Nature Methods* **6,** 451–457 (2009).
34. Luke, S. *The ECJ Owner's Manual. San Francisco, California, A user manual for the ECJ Evolutionary Computation Library* (Evolutionary Computation Library, 2010).



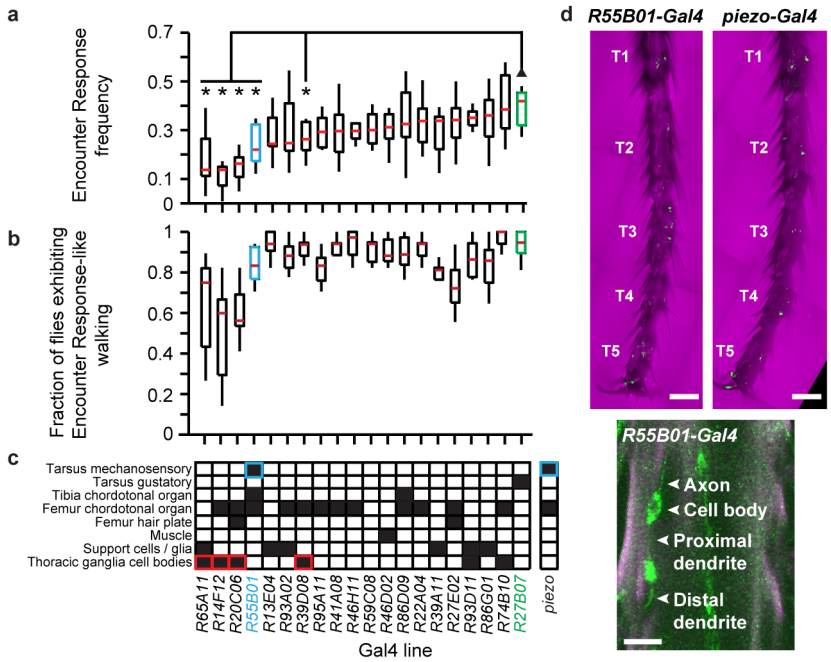
Extended Data Figure 1 | *Drosophila* aggregate and move coherently at high densities. **a**, Images at the start (left) and end (right) of a ~3 h video recording with 100 flies (50 male and 50 female) moving within a large container containing a banana paste dish (left) and an agarose dish (right). **b**, Fly densities on the banana paste dish for each gender or mixture of genders averaged from the 30th through 60th minute of a 90 min experiment ($n = 4$ experiments for each genotype). **c**, The arena for simultaneous odour stimulation and behaviour tracking of *Drosophila* groups. **d**, Laminar flow and odour localization validation using simulated fluid dynamics. High velocity vectors (yellow/red) are present at the odour entry and exit ports while lower, uniform velocity vectors (green/blue) are located within the arena. **e**, A histogram showing the per cent of time avoiding the odour for all flies in all experiments and for each density (colour-coded). Data are from Fig. 1d. **f**, The per cent time avoiding the odour (mean and s.d.) for five different densities of the subset of flies starting in the odour zone that have at some point entered the air zone ($n = 37$, 38, 36, 35, and 38 experiments for 0.06, 0.38, 0.75, 1.13, and 1.5 flies per cm^2 respectively). In contrast to Fig. 1d, the lack of density dependence suggests that flies that leave the odour zone tend not to return. **g**, The formula for a Coherent Motion Index that captures the degree of motion in the same direction (top) and an example of coherent motion away from the odour zone by 9 out of 11 flies total (bottom, cyan). **h**, The Coherent Motion Index for flies in the air (white boxes) or odour (grey boxes) zones during the ten seconds following odour onset. Data are from Fig. 1d. Shown are the results across all tested densities (0.06–1.5 flies per cm^2) for flies that began the experiment in the odour (grey boxes) or the air zone (white boxes). $n = 31$–38 experiments. A single asterisk (*) denotes $P < 0.05$ and a double asterisk (**) denotes $P < 0.01$ for a Bonferroni sign test comparing medians to 0.



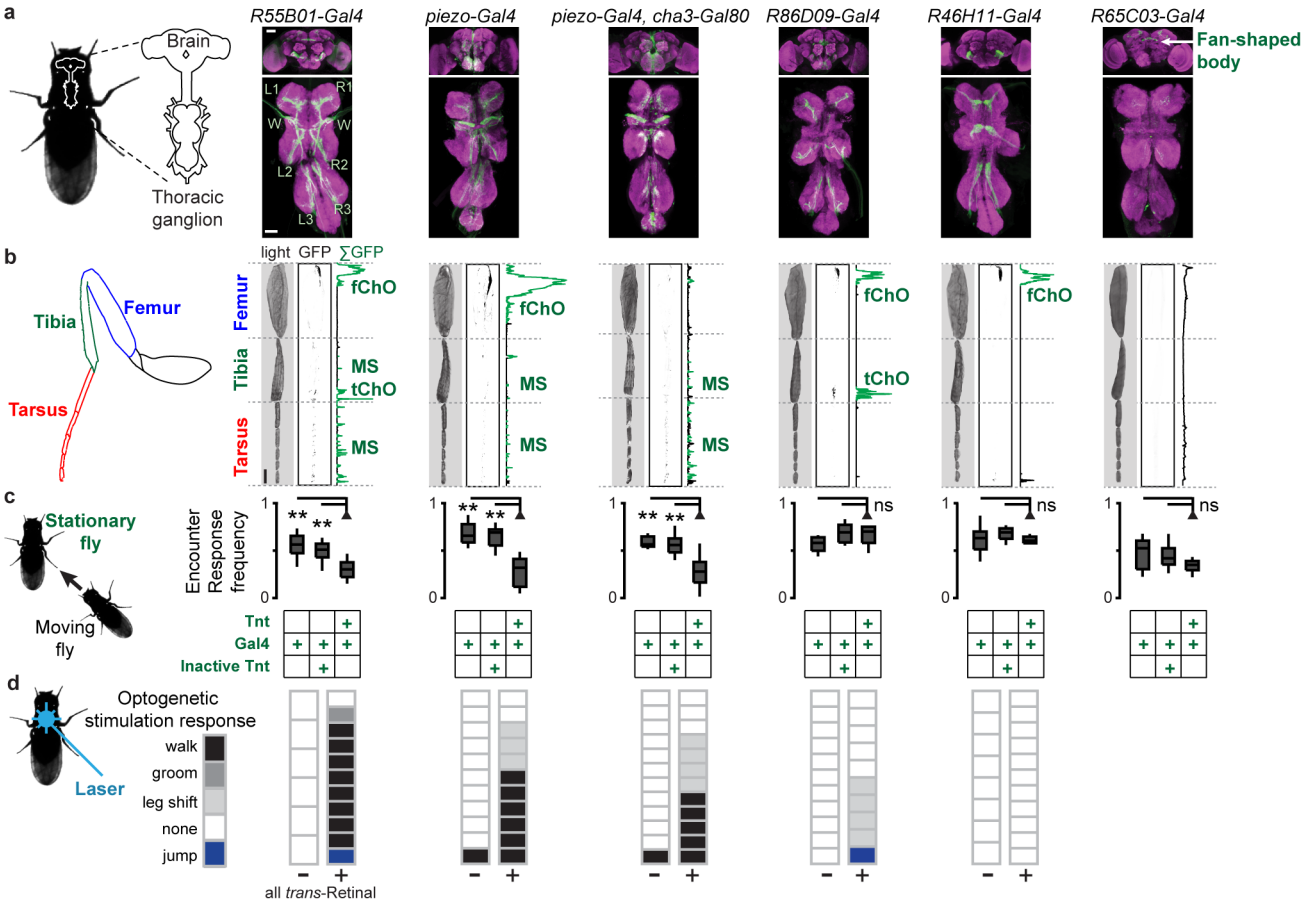
Extended Data Figure 2 | Model parameter determination and the sensitivity of simulated collective behaviour to parameter variation. **a**, **b**, Individual freely walking flies were presented with 5% CO_2 ('odour') or air across the entire arena for 1 min. Mean (solid line) and s.e.m. (translucent shading) walking velocity magnitude (**a**) and forward bout probability (**b**) before, during, and after an odour impulse (black, $n = 45$ flies) or an air impulse control (blue, $n = 43$ flies). Bouts began when velocity exceeded a high threshold of 1 mm s^{-1}. Bouts ended when velocity dipped below a low threshold of 0.5 mm s^{-1}. Short bouts or pauses (<2 frames or 100 ms) were removed by merging the fly's current behavioural state with neighbouring measurements. Grey indicates the period of odour presentation. **c**, Probability for *Drosophila* to turn back when crossing the interface from odour to air and vice versa after a given period of time. Data are from Fig. 1d (density = 0.06). **d**, Scatter plots of *Drosophila* bout lengths during isolation versus Encounter Response bout lengths (red dots) and the double-linear function fitting the data (blue line). $n = 16$ experiments at density = 0.38 flies per cm^2. The graph on the right is a zoom-in of that on the left (dashed box). **e**–**h**, Sensitivity of simulated collective behaviour to $P(bout_{air})$ ranging from Probability = 0 (blue, never initiating spontaneous walking in air) to Probability = 1 (yellow, always initiating spontaneous walking in air) (**e**), $P(bout_{odour})$ ranging from Probability = 0 (blue, never initiating spontaneous walking in odour) to Probability = 1 (yellow, always initiating spontaneous walking in odour) (**f**), P(turn around from air) ranging from Probability = 0 (never turning around from the air zone, blue) to Probability = 1 (always turning around from the air zone, yellow) (**g**), P(turn away from odour) ranging from Probability = 0 (never turning around from the odour zone, blue) to Probability = 1 (always turning around from the odour zone, yellow) (**h**). In all panels, each coloured line indicates the mean per cent time avoiding the odour across densities, the black line indicates the simulation result for parameter values taken from real fly data, $n = 10{,}902$ for all data-points, and superimposed are the mean values for real flies (black circles).



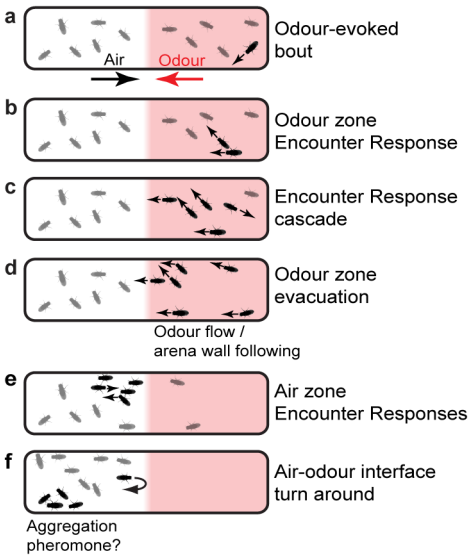
Extended Data Figure 3 | Encounter Response kinematics for inter-fly or metallic disc touches. **a**, Schematic of octant colour-coding. Each Encounter Response trajectory is assigned to the perimeter octant bisected by a line drawn to the nearest neighbouring fly during an Encounter. A head octant (red) is included here but these responses likely represent front leg touches. **b**, The mean (solid lines) and standard error (translucent areas) for Encounter Response trajectories (right) colour-coded by the relative location of the neighbouring fly as in panel **a**. The scale bar is 1 mm. **c**, Boxplot of mean forward (top), sideways (middle), and angular (bottom) velocities for the first 0.5 s of Encounter Responses (n = 112–244 Encounters with duration >0.5 s) in the olfactory avoidance experiment from Fig. 1d (density = 0.75 flies per cm^2). Velocities are colour-coded by octant. **d**, Schematic of touch-point colour-coding for high-resolution inter-fly touch response experiments. Each walking trajectory is colour-coded by the appendage touched by a neighbouring fly. Data are from Fig. 3c. **e**, Boxplot of mean forward (top), sideways (middle), and angular (bottom) velocities for the first 0.16 s of touch responses. Velocities are colour-coded by touch-point. **f**, Schematic of touch-point colour-coding for mechanical touch response experiments. Each touch response trajectory is assigned to the appendage touched by a metallic disc. Data are from Fig. 3d. **g**, Boxplot of mean forward (top), sideways (middle), and angular (bottom) velocities for the first 0.5 s of touch responses. Velocities are colour-coded by touch-point.



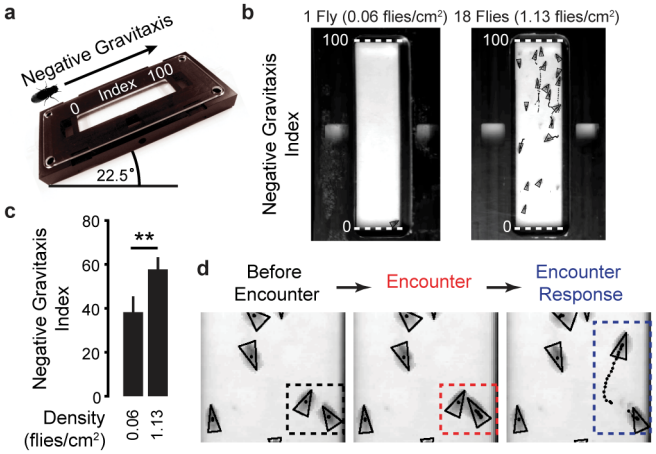
Extended Data Figure 4 | A behavioural screen for neurons mediating Encounter Responses and their leg expression patterns. **a**, Frequency of Encounter Responses for each Gal4 driver expressing *UAS-Tnt*. Driver lines are sorted by median frequency of Encounter Responses. A single asterisk (*) indicates $P < 0.05$ for a Bonferroni-corrected Mann–Whitney *U*-test comparing a given line against a gustatory neuron expression line, *R27B07-Gal4* (green). Density = 1.13 flies per cm^2 and $n = 10$ experiments for each line. The selected line, *R55B01-Gal4*, drives expression in distal leg mechanosensory neurons (cyan). **b**, The fraction of flies in each experiment exhibiting walking velocities that meet the criteria for Encounter Responses (mean velocity magnitude greater than 5 mm s^{-1} for more than 0.5 s) at any time during the experiment. Lines are sorted and colour-coded as in panel **a**. **c**, The identity and leg expression patterns of Gal4 drivers tested in the screen. Black boxes denote the presence of a given cell class. A cyan outline indicates distal leg mechanosensory neuron expression. A red outline indicates thoracic ganglion expression in lines with significant reductions in Encounter Response frequency. The expression pattern is also shown for *piezo-Gal4*, which was used in subsequent experiments to refine identification of the leg mechanosensory neuron class required for Encounter Responses. **d**, Tarsal segments for *w*;*UAS-CD4:tdGFP*;*R55B01-Gal4* (left) and *w*;*UAS-CD4:tdGFP*;*piezo-Gal4* (right) flies. Each tarsal segment is labelled from proximal to distal (T1-T5). Endogenous GFP fluorescence (green) is superimposed upon a transmitted light image (magenta). The scale bars are 30 µm. Below is a high-resolution image of a mechanosensory sensilla neuron on the tarsus of a *w*;*UAS-CD4:tdGFP*;*R55B01-Gal4* fly. Endogenous GFP fluorescence (green) is superimposed on cuticular autofluorescence (magenta). The axon, cell body, and dendrite of this neuron are labelled. The scale bar is 10 µm.



Extended Data Figure 5 | Leg mechanosensory sensilla neurons, but not chordotonal organs, are necessary and sufficient for Encounter Responses. We identified five lines expressing Gal4 in different subsets of mechanosensory neurons (*R55B01-Gal4*, *piezo-Gal4*, *piezo-Gal4;cha3-Gal80, R86D09-Gal4*, and *R46H11-Gal4*) and one line expressing Gal4 in the fan-shaped body (*R65C03-Gal4*) as a control for fan-shaped body expression in *R55B01-Gal4*. **a**, Brain and thoracic ganglion expression for Gal4 lines driving *UAS-CD4:tdGFP*. Immunostaining is shown for the neuropil marker nc82 (magenta) and CD4:tdGFP (green). Sensory neuron projections from the wings ('W') and legs (R1–R3 and L1–L3) are labelled for *R55B01-Gal4*. Importantly, neurons expressing GFP in the brains of *R55B01-Gal4* and *piezo-Gal4*; *cha3-Gal80* flies are different, implying that they are not responsible for the production of Encounter Responses. The scale bars are 40 μm. **b**, Transmitted light images, inverted GFP fluorescence images (GFP indicated in black), and summed fluorescence of Gal4 driver legs expressing CD4:tdGFP. Autofluorescent cuticle and pretarsus debris are indicated in black. GFP expression is shown in green. When present, the femoral chordotonal organ ('fChO'), tibial chordotonal organ ('tChO') and mechanosensory sensilla neurons ('MS') are labelled. The scale bar is 100 μm. **c**, The frequency of Encounter Responses for a parental control ('Gal4'), Gal4 line neurons expressing an inactive tetanus toxin control ('Gal4' and 'Inactive Tnt'), or Gal4 line neurons expressing tetanus toxin ('Gal4' and 'Tnt'). $n = 10–15$ experiments for each condition. **d**, Blue laser pulse stimulation responses of Gal4 line flies expressing *UAS-ChR2* in the absence (left) or presence (right) of the essential cofactor all *trans*-Retinal ($n = 6–12$ flies for each condition). Each box indicates the response for a single fly ('walk', 'groom', 'leg shift', 'none', or 'jump').



Extended Data Figure 6 | Schematic of collective odour avoidance in *Drosophila*. **a**, A group of flies experiences odour flow on the right half of the arena. The direction of odour or air flow is indicated by red and black arrows, respectively. Odour increases the probability of spontaneous walking (black fly). **b**, Walking increases the probability of encountering a stationary fly, producing an Encounter Response. **c**, Walking flies cause additional Encounters and a cascade of Encounter Responses in the odour zone. **d**, Walking flies pass into the non-odour zone through interactions with the arena walls and possibly by sensing the direction of odour flow. **e**, The influx of walking flies to the air zone results in additional Encounter Responses. **f**, The propensity to turn around at the air–odour interface (perhaps compounded by the effects of unknown aggregation pheromones) causes flies to remain in the air zone, resulting in odour avoidance.



Extended Data Figure 7 | Collective negative gravitaxis in *Drosophila*. **a**, A schematic of the negative gravitaxis experiment. Flies are placed at the lowest point of a behavioural arena tilted at 22.5°. The flies' positions are normalized to the long-axis of the arena ranging from 0 (arena bottom, lowest elevation) to 100 (arena top, highest elevation). **b**, Image of flies (black triangles) and their trajectories during 1 s (black dotted lines) in the negative gravitaxis experiment. Shown are representative images of an experiment with one fly (density = 0.06 flies per cm^2) and an experiment with 18 flies (density = 1.13 flies per cm^2). Negative Gravitaxis Index value positions of 0 (lowest elevation in the arena) and 100 (highest elevation in the arena) are shown (white-dashed lines). **c**, To obtain a Negative Gravitaxis Index for a given fly, its position was averaged during the second minute of the experiment. Shown are the mean and s.d. of Negative Gravitaxis Indices for wild-type animals at densities of either 0.06 or 1.13 flies per cm^2 (n = 28 and 30 experiments, respectively). **d**, Images of two flies (left, black triangles in black dashed box) undergoing an Encounter (middle, red dashed box) that results in an Encounter Response (right, blue dashed box) during a negative gravitaxis experiment.

Extended Data Table 1 | *P* values for data in main figures

Figure	Comparison	*P* value (uncorrected)	Number of Comparisons
1d	6 vs. 1 fly	3.37×10^{-6}	4
	12 vs. 1 fly	4.20×10^{-13}	
	18 vs. 1 fly	4.96×10^{-21}	
	24 vs. 1 fly	6.77×10^{-18}	
2d	12 vs. 6 flies	1.67×10^{-11}	1
	18 vs. 12 flies	3.91×10^{-11}	1
	24 vs. 18 flies	4.50×10^{-11}	1
2e	6 flies, Enc. vs. Iso.	3.18×10^{-4}	1
	12 flies, Enc vs. Iso.	3.78×10^{-6}	1
	18 flies, Enc vs. Iso.	1.30×10^{-3}	1
2g	6 vs. 1 fly	3.69×10^{-8}	4
	12 vs. 1 fly	1.87×10^{-17}	
	18 vs. 1 fly	4.99×10^{-19}	
	24 vs. 1 fly	2.27×10^{-16}	
3a	Light vs. Dark	1.73×10^{-2}	5
	Anosmic vs. Dark	4.27×10^{-1}	
	nanchung vs. Dark	1.73×10^{-2}	
	piezo vs. Dark	1.13×10^{-2}	
	nompC vs. Dark	1.00×10^{-3}	
3e	oe+ vs. oe-	5.55×10^{-1}	1
3g	Tnt vs. Gal4>Tnt	1.28×10^{-4}	3
	Gal4 vs. Gal4>Tnt	2.74×10^{-4}	
	Inactive Tnt vs. Gal4>Tnt	1.10×10^{-3}	
4a	Inactive Tnt, 18 vs. 1 fly	2.99×10^{-6}	1
	Tnt, 18 vs. 1 fly	5.54×10^{-1}	1
4b	Inactive Tnt, 18 vs. 1 fly	3.90×10^{-8}	1
	Tnt, 18 vs. 1 fly	3.75×10^{-2}	1
4c	$nompC^{+/-}$, 18 vs. 1 fly	1.12×10^{-5}	1
	$nompC^{-/-}$, 18 vs. 1 fly	1.62×10^{-1}	1
4d	3 vs. 1 fly	3.98×10^{-2}	3
	6 vs. 1 fly	2.98×10^{-9}	
	12 vs. 1 fly	3.12×10^{-25}	
4e	3 vs. 1 fly	8.87×10^{-18}	3
	6 vs. 1 fly	1.76×10^{-18}	
	12 vs. 1 fly	3.60×10^{-38}	

The uncorrected *P* values for each main figure panel and its associated comparison are indicated. The number of comparisons used for post-hoc Bonferroni correction for multiple comparisons is also shown.

Extended Data Table 2 | *P* values for data in Extended Data figures

Figure	Comparison	*P* value (uncorrected)	Number of Comparisons
1h	Air vs. Odour, 6 flies	8.90×10^{-1}	1
	Air vs. Odour, 12 flies	2.90×10^{-3}	1
	Air vs. Odour, 18 flies	2.16×10^{-4}	1
	Air vs. Odour, 24 flies	1.40×10^{-3}	1
4a	*R65A11* vs. *R27B07*	5.01×10^{-4}	19
	R14F12 vs. *R27B07*	2.18×10^{-4}	
	R20C06 vs. *R27B07*	8.15×10^{-5}	
	R55B01 vs. *R27B07*	2.50×10^{-3}	
	R13E04 vs. *R27B07*	5.10×10^{-3}	
	R93A02 vs. *R27B07*	5.69×10^{-2}	
	R39D08 vs. *R27B07*	2.50×10^{-3}	
	R95A11 vs. *R27B07*	1.26×10^{-2}	
	R41A08 vs. *R27B07*	7.94×10^{-2}	
	R46H11 vs. *R27B07*	1.78×10^{-2}	
	R59C08 vs. *R27B07*	1.81×10^{-2}	
	R46D02 vs. *R27B07*	4.02×10^{-2}	
	R86D09 vs. *R27B07*	3.25×10^{-1}	
	R22A04 vs. *R27B07*	2.37×10^{-1}	
	R39A11 vs. *R27B07*	4.18×10^{-2}	
	R27E02 vs. *R27B07*	2.37×10^{-1}	
	R93D11 vs. *R27B07*	2.18×10^{-1}	
	R86G01 vs. *R27B07*	3.72×10^{-1}	
	R74B10 vs. *R27B07*	8.44×10^{-1}	
5c	*R55B01-Gal4*		2
	Gal4 vs.Gal4>Tnt	1.10×10^{-3}	
	Gal4>Inactive vs. Gal4>Tnt	1.70×10^{-3}	
	piezo-Gal4		2
	Gal4 vs.Gal4>Tnt	1.24×10^{-4}	
	Gal4>Inactive vs. Gal4>Tnt	1.50×10^{-4}	
	piezo-Gal4, cha3-Gal80		2
	Gal4 vs.Gal4>Tnt	2.74×10^{-4}	
	Gal4>Inactive vs. Gal4>Tnt	1.10×10^{-3}	
	R86D09-Gal4		2
	Gal4 vs.Gal4>Tnt	2.07×10^{-1}	
	Gal4>Inactive vs. Gal4>Tnt	6.94×10^{-1}	
	R46H11-Gal4		2
	Gal4 vs.Gal4>Tnt	6.89×10^{-1}	
	Gal4>Inactive vs. Gal4>Tnt	3.51×10^{-2}	
	R65C03-Gal4		2
	Gal4 vs.Gal4>Tnt	7.94×10^{-2}	
	Gal4>Inactive vs. Gal4>Tnt	2.62×10^{-2}	
7c	18 vs. 1 fly	3.63×10^{-11}	1

The uncorrected *P* values for each Extended Data figure panel and its associated comparison are indicated. The number of comparisons used for post-hoc Bonferroni correction for multiple comparisons is also shown.

LETTER

doi:10.1038/nature14022

Identification of a mast-cell-specific receptor crucial for pseudo-allergic drug reactions

Benjamin D. McNeil[1], Priyanka Pundir[2], Sonya Meeker[3], Liang Han[1], Bradley J. Undem[3], Marianna Kulka[2,4] & Xinzhong Dong[1,5]

Mast cells are primary effectors in allergic reactions, and may have important roles in disease by secreting histamine and various inflammatory and immunomodulatory substances[1,2]. Although they are classically activated by immunoglobulin (Ig)E antibodies, a unique property of mast cells is their antibody-independent responsiveness to a range of cationic substances, collectively called basic secretagogues, including inflammatory peptides and drugs associated with allergic-type reactions[1,3]. The pathogenic roles of these substances have prompted a decades-long search for their receptor(s). Here we report that basic secretagogues activate mouse mast cells *in vitro* and *in vivo* through a single receptor, Mrgprb2, the orthologue of the human G-protein-coupled receptor MRGPRX2. Secretagogue-induced histamine release, inflammation and airway contraction are abolished in Mrgprb2-null mutant mice. Furthermore, we show that most classes of US Food and Drug Administration (FDA)-approved peptidergic drugs associated with allergic-type injection-site reactions also activate Mrgprb2 and MRGPRX2, and that injection-site inflammation is absent in mutant mice. Finally, we determine that Mrgprb2 and MRGPRX2 are targets of many small-molecule drugs associated with systemic pseudo-allergic, or anaphylactoid, reactions; we show that drug-induced symptoms of anaphylactoid responses are significantly reduced in knockout mice; and we identify a common chemical motif in several of these molecules that may help predict side effects of other compounds. These discoveries introduce a mouse model to study mast cell activation by basic secretagogues and identify MRGPRX2 as a potential therapeutic target to reduce a subset of drug-induced adverse effects.

Responsiveness to basic secretagogues is conserved among mammals[4] and is also found in birds[5], indicating an ancient, fundamental role for its mechanism. Many basic secretagogues are endogenous peptides, often linked to inflammation; however, they activate connective tissue mast cells only at high concentrations and independent of their canonical receptors, so another mechanism of stimulation must exist[6]. Several candidate proteins that bind polycationic compounds have been proposed as basic secretagogue receptors[6–9]. Among these, MRGPRX2 has been screened with the most compounds[8,10–14], and short interfering RNA (siRNA) knockdown studies support at least a partial role for *MRGPRX2* in activation by four non-canonical basic secretagogues[11,13]. However, no direct *in vivo* study or knockout model has been employed for any candidate. The investigation of MRGPRX2 in mice is complicated because

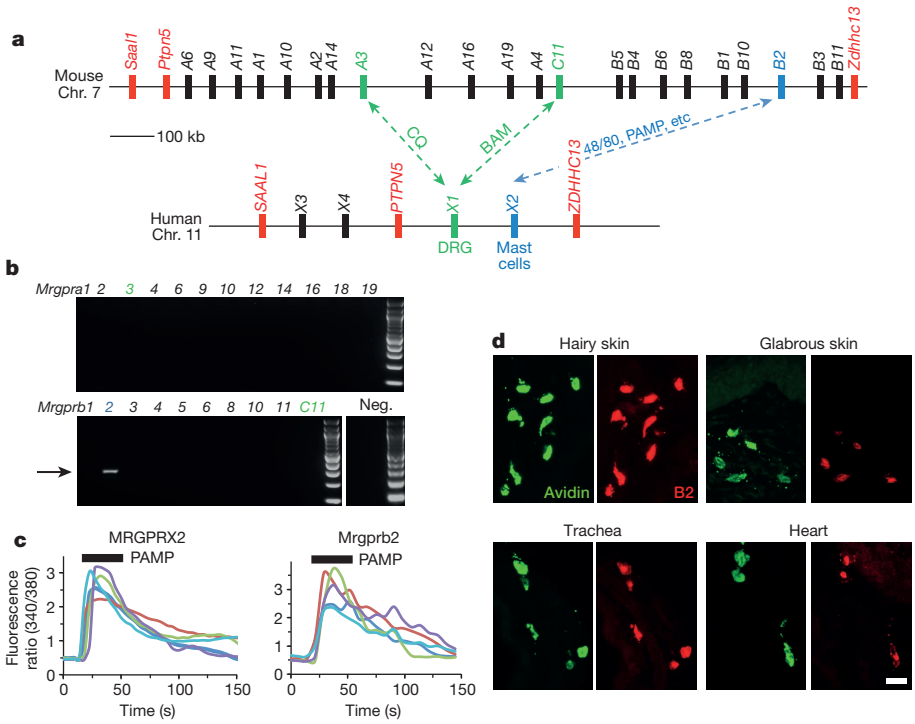


Figure 1 | Mrgprb2 is the orthologue of human MRGPRX2. a, Diagram of mouse and human *Mrgpr* genomic loci. Mouse *Mrgpra3* and *Mrgprc11* are orthologues of human *MRGPRX1*, determined by expression and ligand specificity[15]. The MRGPRX2 orthologue Mrgprb2 is described in this study. Chr., chromosome. **b**, Results from a stringent RT–PCR screen identifying *Mrgprb2* transcript (arrow) in mouse peritoneal mast cells. The negative control (Neg.) omitted reverse transcriptase. RT–PCR for *Mrgprb2* was repeated at least four times. **c**, Example traces of intracellular calcium concentrations $[Ca^{2+}]_i$, measured by ratiometric Fura-2 imaging, from Mrgprb2-HEK or MRGPRX2-HEK cells exposed to 20 μM PAMP(9–20) (duration indicated by black line). Each trace is a response from a unique cell. **d**, Representative confocal images from BAC transgenic mouse tissues in which tdTomato expression is controlled by enhanced green fluorescent protein (eGFP)-Cre expression from the *Mrgprb2* locus (see Methods). Avidin staining was used to identify mast cells. Percentages of avidin-positive mast cells that were also tdTomato positive: glabrous skin, 97.5%; hairy skin, 90.1%; trachea, 97.2%; heart, 87.1%. Percentages of tdTomato-positive cells that were also avidin positive: glabrous skin, 99.2%; hairy skin, 100%; trachea, 98.3%; heart, 99%. $n = 3$ mice and >300 cells counted per tissue, except $n = 2$ and >100 cells counted in the heart. Scale bar, 20 μm.

[1]The Solomon H. Snyder Department of Neuroscience, Department of Neurosurgery, Center for Sensory Biology, Johns Hopkins University, School of Medicine, Baltimore, Maryland 21205, USA. [2]Department of Medical Microbiology and Immunology, University of Alberta, Edmonton, Alberta T6G 2E1, Canada. [3]Department of Medicine, Division of Allergy and Clinical Immunology, Johns Hopkins University, School of Medicine, Baltimore, Maryland 21205, USA. [4]National Institute for Nanotechnology, National Research Council Canada, Edmonton, Alberta T6G 2M9, Canada. [5]Howard Hughes Medical Institute, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA.

the gene cluster containing the four human *MRGPRX* members is dramatically expanded in mice, consisting of 22 potential coding genes, many with comparable sequence identity to *MRGPRX2* (Fig. 1a). Therefore, a mouse *MRGPRX2* orthologue must be determined by expression pattern and pharmacology. A stringent polymerase chain reaction with reverse transcription (RT–PCR) screen in mouse primary mast cells uncovered a band for a single family member, *Mrgprb2* (Fig. 1b), whereas *MRGPRX1* orthologues were not expressed at relevant levels (Extended Data Fig. 1a, b). Functionally, HEK293 cells heterologously expressing mouse Mrgprb2 (Mrgprb2-HEK) responded to the MRGPRX2 agonist proadrenomedullin amino-terminal 20 peptide, fragment 9–20 (PAMP(9–20))[14] (Fig. 1c) and compound 48/80 (48/80), a classical mast cell activator and canonical basic secretagogue (Extended Data Fig. 2). Mrgprb2-HEK cells also responded to other MRGPRX2 ligands, including the basic secretagogue Substance P, but had no response to the MRGPRX1 ligand chloroquine[15]; no closely related family members in mice responded to any compound (Extended Data Figs 1c and 2a, c). To determine the expression of *Mrgprb2*, we generated *Mrgprb2* bacterial artificial chromosome (BAC) transgenic mice in which the expression of *eGFP-Cre* recombinase was under the control of the *Mrgprb2* promoter. Strikingly, Cre expression patterns indicate that *Mrgprb2* expression is highly specific to connective tissue mast cells (Fig. 1d and Extended Data Figs 3, 4). Together, the pharmacological and expression data indicate that Mrgprb2 is the mouse orthologue of human MRGPRX2.

Next, we determined whether Mrgprb2 is the basic secretagogue receptor in mouse mast cells. The *Mrgprb2* genomic locus contains too much repetitive sequence to permit gene targeting through homologous recombination (Extended Data Fig. 5a). Therefore, we used a zinc-finger-nuclease-based strategy to generate a mouse line with a 4 base pair (bp) deletion in the *Mrgprb2* coding region (Mrgprb2^{MUT} mice), resulting in a frameshift mutation and early termination shortly after the first transmembrane domain (Extended Data Fig. 5b–d). The mutation was stable and inheritable (Extended Data Fig. 5c), so we regard Mrgprb2^{MUT} as a functional null. Mast cell numbers were comparable in tissues of wild-type and Mrgprb2^{MUT} mice, indicating that Mrgprb2 is not essential for mast cell survival or targeting to tissue (Extended Data Fig. 6a). The responsiveness of peritoneal mast cells to anti-IgE antibodies (Fig. 2a) and endothelin (Extended Data Fig. 7) was also comparable, demonstrating that Mrgprb2 mutation does not globally impair IgE or G-protein-coupled receptor (GPCR)-mediated mast cell signalling. However, 48/80-induced mast cell activation (Fig. 2a) and tissue histamine release were essentially abolished in mutant mast cells (Fig. 2b and Extended Data Fig. 6b). Furthermore, we found that 48/80-evoked tracheal contraction (Fig. 2c) and hindpaw inflammation (extravasation and swelling; Fig. 2d) were almost completely absent in an Mrgprb2^{MUT} background, while antigen (Fig. 2c) and anti-IgE evoked responses (Extended Data Fig. 8) were comparable to wild-type mice. Finally, we found that four additional basic secretagogues, as well as the MRGPRX2 agonists PAMP(9–20) and cortistatin[10], strongly activated wild-type but not Mrgprb2^{MUT} mast cells (Fig. 2e and Extended Data Fig. 9a). HEK293 cells expressing Mrgprb2 or MRGPRX2 (MRGPRX2-HEK) also responded to these secretagogues (Extended Data Fig. 2). Taken together, we conclude that Mrgprb2 is the mouse mast cell basic secretagogue receptor. It is likely that the list of small, basic peptides that activate Mrgprb2 is greater than the number in this study; indeed, dozens of such peptides have been shown to activate mast cells[3,6,16,17]. Notably, human MRGPRX2 is much more sensitive to substance P than mouse Mrgprb2 (Extended Data Fig. 2c), suggesting a potential species-specific role for substance P in mast cell signalling.

We next considered whether Mrgprb2 factors in allergic-type reactions. We specifically addressed drug-induced reactions because many therapeutic drugs are cationic. Up to 15% of drug-induced adverse reactions appear to be allergic in nature; however, many do not correlate well with IgE antibody titre, indicating that antibody-independent, or pseudo-allergic, mechanisms participate[18]. We focused first on peptidergic drugs

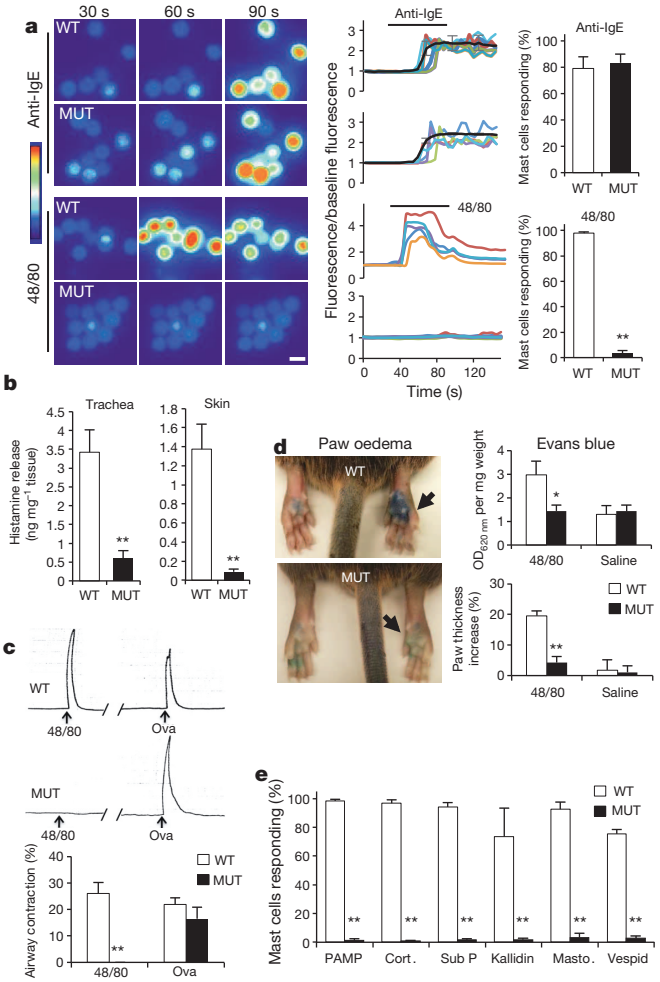


Figure 2 | Mrgprb2 is the mouse mast cell basic secretagogue receptor. **a**, Left, representative Fluo-4 fluorescence heat map images of mouse peritoneal mast cells showing changes in $[Ca^{2+}]_i$ induced by bath application of anti-IgE (5 μg ml^{-1}) or 48/80 (10 μg ml^{-1}). Middle, representative imaging traces. Each colour line represents an individual cell. Black lines in 'anti-IgE' panels are average traces for each genotype. Note that $[Ca^{2+}]_i$ traces are similar between wild-type (WT) and mutant (MUT) groups. Right, quantification of responding cells ($n = 3$ per genotype; >150 cells counted per condition). Anti-IgE responses were not significantly different. Scale bar, 10 μm. **b**, Histamine release into the supernatant from trachea and abdominal skin from wild-type and Mrgprb2^{MUT} mice after exposure to 48/80 (30 μg ml^{-1}) for 30 min at 37 °C. $n = 5$ for trachea, $n = 8$ for skin. **c**, Top, representative traces showing contractions of trachea isolated from wild-type and Mrgprb2^{MUT} mice (previously sensitized to ovalbumin (Ova)) in response to 48/80 (30 μg ml^{-1}) or ovalbumin (10 μg ml^{-1}; that is, IgE dependent). Bottom, average data; maximum total contraction determined as response to 10 μM carbamycholine added at the end of the experiment. $n = 5$ for 48/80 wild type, $n = 3$ for 48/80 MrgprB2^{MUT}. **d**, Left, representative images of Evans blue stained extravasation 15 min after intraplantar injection of 48/80 (right, arrow, 10 μg ml^{-1}, 5 μl in saline) or saline (left). Right, quantification of Evans blue leakage into the paw and paw thickness increase after 15 min. $OD_{620\,nm}$, optical density at 620 nm. *$P < 0.02$ ($n = 5$ wild type, $n = 6$ MrgprB2^{MUT}). Differences after saline injection were not significant. **e**, Quantification of wild-type and Mrgprb2^{MUT} mast cell responsiveness to MRGPRX2 ligands and basic secretagogues, assayed using Fluo-4 imaging. Concentrations of substances (in μM): PAMP(9–20), 20; cortistatin-14 (Cort.), 20; substance P (Sub P), 200; kallidin, 200; mastoparan (masto.; a component of wasp venom), 20; vespid mastoparan, 20. $n = 3$ per genotype; >150 cells counted per secretagogue. Data are presented as mean ± standard error of the mean (s.e.m.). Two-tailed unpaired Student's *t*-test was used to determine significance in statistical comparisons, and differences were considered significant at $P < 0.05$. *$P < 0.05$, **$P < 0.01$.

because most are introduced subcutaneously or intramuscularly at millimolar concentrations (Supplementary Information), high enough for cationic peptides to activate mast cells. The most frequent allergic-type response described in the FDA labels of these drugs is an injection-site reaction (ISR), a local swelling and/or flare of variable size that can be accompanied by pain or pruritus. In a survey of FDA-approved peptidergic drugs, we found that the vast majority associated with ISRs are cationic (Supplementary Information). We found that representative members of all common, commercially available classes of these cationic drugs activated mast cells in an Mrgprb2-dependent manner, whereas the innocuous protein insulin had no effect (Fig. 3a and Extended Data Fig. 9b, c). Consistently, all of these peptides except insulin activate both Mrgprb2-HEK and MRGPRX2-HEK cells (Extended Data Fig. 2). We selected the drug icatibant for further study because it induces ISRs in nearly every patient[19]. Icatibant at the clinical concentration induced extensive extravasation and swelling, similar to human ISRs, in wild-type mice but not in Mrgprb2^{MUT} mice (Fig. 3b). Mice pretreated with the mast cell stabilizer ketotifen also showed no inflammation (without ketotifen: 40.7 ± 2.1% increase in paw thickness; with ketotifen: 3.1 ± 0.6% increase; $n = 4$ each; $P = 2.2 \times 10^{-6}$), strongly indicating that mast cells mediated the inflammation. Furthermore, icatibant (as well as positive controls 48/80 and mastoparan) induced histamine release from wild-type peritoneal mast cells, whereas Mrgprb2^{MUT} mast cells released substantially less histamine (Fig. 3c). However, IgE-mediated histamine release was unaffected by *Mrgprb2* deletion (Fig. 3c). These data lead us to anticipate that drug-induced ISRs may be alleviated by targeting MRGPRX2 or by using peptides with less potent MRGPRX2 agonist properties.

Next, we explored the possibility that Mrgprb2 mediates pseudo-allergic reactions induced by small molecules. We focused on intravenous drugs because they are often administered rapidly and in high doses, and thus are more likely to achieve high blood concentrations and rapid tissue distribution than drugs administered through other routes. Symptoms of pseudo-allergic reactions after intravenous administration, which at their most severe are called anaphylactoid, include skin flushing or rash, changes in blood pressure or heart rate, and bronchospasms[20]. We based our initial search on the structure of 48/80. While the structure–function relationship of 48/80 as an MRGPRX2 agonist is unknown, a cyclized variant containing a tetrahydroisoquinoline (THIQ) motif (Fig. 4a) is reported to be seven times more potent than 48/80 as a mast cell degranulator[21]. A search of FDA-approved drugs containing a THIQ recovered members of the nicotinic receptor antagonist non-steroidal neuromuscular blocking drugs (NMBDs), including tubocurarine and atracurium (Fig. 4b). NMBDs are used routinely in surgery to reduce unwanted muscle movement and allow intratracheal intubation for mechanical ventilation. Intriguingly, NMBDs alone are responsible for nearly 60% of allergic reactions in a surgical setting[22], and all except succinylcholine induce histamine release in humans[23]. We found that members of all NMBD families (Supplementary Information) except succinylcholine activated mast cells in an Mrgprb2-dependent manner at concentrations as low as 0.5% of the clinical injection concentration (Fig. 4c and Extended Data Fig. 9d). Interestingly, rocuronium does not contain a THIQ but has a bulky hydrophobic group with a charged nitrogen within several angstroms (Fig. 4b), reminiscent of 48/80. Therefore, we searched using modifications of the THIQ motif and the 48/80 structure, including changes in cyclization and position of the positive or polar nitrogen, limiting our assay to intravenous drugs at high injection concentrations. We identified the fluoroquinolone family of antibiotics as having a similar motif (Fig. 4d). Like NMBDs, these are associated with allergic-type reactions[24,25] and can activate mast cells[26,27]. We found that the four members approved for intravenous use activated Mrgprb2-HEK and MRGPRX2-HEK cells (Extended Data Fig. 2), and

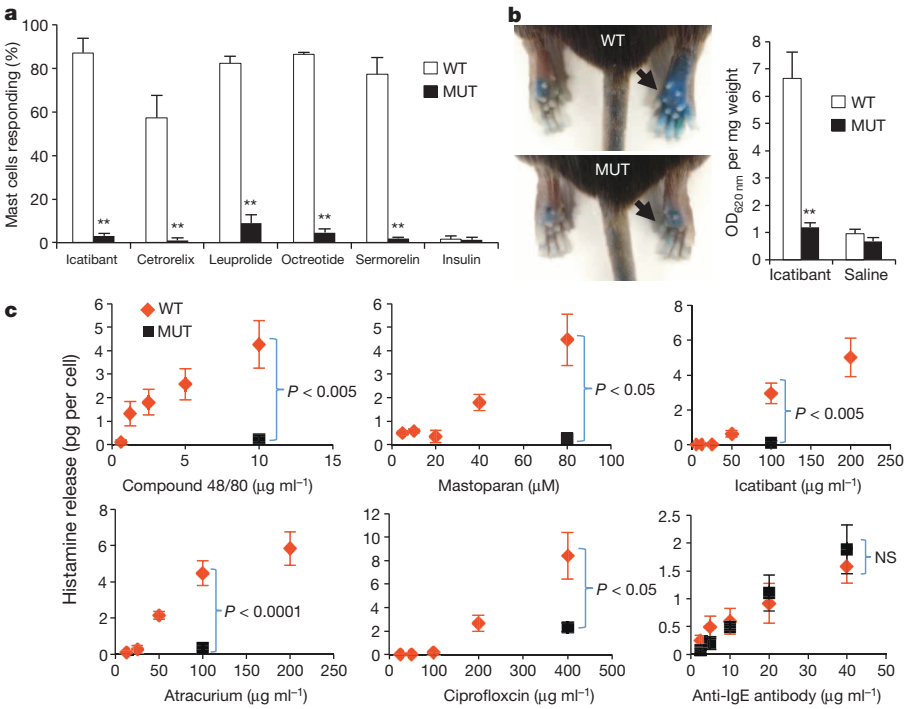


Figure 3 | Mrgprb2 mediates mast cell responsiveness and side effects of peptidergic therapeutic drugs. **a**, Percentage of responding cells from wild-type (WT) and Mrgprb2^{MUT} (MUT) peritoneal mast cells after drug application, assayed using Fluo-4 imaging. Concentrations of drugs (in μg ml^{-1}): icatibant, 50; cetrorelix, 20; leuprolide, 100; octreotide, 10; sermorelin, 60; insulin, 80. $n = 3$ per genotype; >150 cells counted per substance, except >100 cells counted for insulin. Difference between insulin responsiveness was not significant. **b**, Left, representative images of Evans blue stained extravasation 15 min after intraplantar injection of icatibant (right, arrow, 10 mg ml^{-1}, 5 μl in saline) or saline (left). Right, quantification of Evans blue leakage into the paw after 15 min. $n = 6$ per genotype. Difference after saline injection was not significant. **c**, Total histamine release from wild-type (red diamonds) and Mrgprb2^{MUT} (black squares) mice after incubation with named substances. Note that no significant difference between wild-type and Mrgprb2^{MUT} cells was found at any dose of anti-IgE antibody. Experiments were repeated >3 times. Data are presented as mean ± s.e.m. Two-tailed unpaired Student's *t*-test: $*P < 0.05$, $**P < 0.01$. NS, not significant.

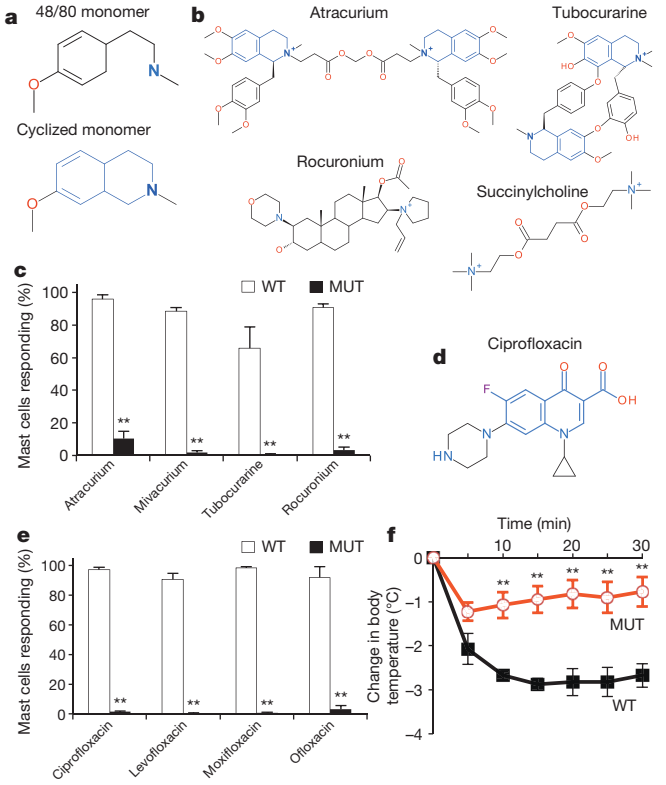


Figure 4 | Mrgprb2 mediates mast cell responsiveness and side effects of small-molecule therapeutic drugs. **a**, Structures of 48/80 and a cyclized variant. The THIQ motif is highlighted in blue. **b**, Structures of representative members of all NMBD classes (see Supplementary Information). THIQ motifs are highlighted in blue. Note that only succinylcholine lacks a bulky hydrophobic group. **c**, Percentage of responding cells from wild-type (WT) and Mrgprb2^{MUT} (MUT) peritoneal mast cells after application of various NMBDs, assayed using Fluo-4 imaging. Concentrations of drugs (in $\mu g\,ml^{-1}$): atracurium, 50; mivacurium, 20; tubocurarine, 30; rocuronium, 500. $n = 3$ mice per genotype; >150 cells counted per substance. **d**, Structure of ciprofloxacin, with the motif common to all fluoroquinolones highlighted in blue. Note the nitrogens close to the quinolone motif. **e**, Percentage of responding cells from wild-type and Mrgprb2^{MUT} peritoneal mast cells after fluoroquinolone application, assayed using Fluo-4 imaging. Concentrations of drugs (in $\mu g\,ml^{-1}$): ciprofloxacin, 200; levofloxacin, 500; moxifloxacin, 160; ofloxacin, 400. $n = 3$ mice per genotype; >150 cells counted per substance. **f**, Changes in body temperature after intravenous injection of ciprofloxacin (1.5 mg in 125 µl saline) at time 0. $n = 4$ mice per genotype. Data are presented as mean ± s.e.m. Two-tailed unpaired Student's *t*-test: $*P < 0.05$, $**P < 0.01$.

mast cells in an Mrgprb2-dependent manner (Fig. 4e and Extended Data Fig. 9e). Correspondingly, atracurium and ciprofloxacin induced histamine release in wild-type peritoneal mast cells and substantially less so in Mrgprb2^{MUT} mast cells (Fig. 3c). We selected ciprofloxacin for *in vivo* tests of anaphylaxis, which in mice is measured most often by a drop in body temperature, probably due to changes in blood pressure and peripheral vasodilation[28]. Rodents are nearly immune to histamine toxicity at a systemic level, contrary to other experimental organisms[4], but can be rendered sensitive to mast cell activators and secreted products by pretreatment with β-adrenergic blockers[29,30]. Under these conditions, a high dose of ciprofloxacin induced a rapid drop in body temperature that was very slow to recover, while Mrgprb2^{MUT} mice showed a much smaller drop that recovered quickly (Fig. 4f). These results establish that mast cell activation through Mrgprb2 is an off-target effect of fluoroquinolones and other drugs.

Finally, we determined whether drugs associated with pseudo-allergies activate human mast cells through MRGPRX2. We found that representative members of each examined drug class evoked release of histamine, tumour necrosis factor (TNF), prostaglandin D_2 (PGD_2) and β-hexosaminidase from LAD2 cells (Extended Data Fig. 10a). 48/80 and mastoparan were used as positive controls. Importantly, MRGPRX2-siRNA-treated LAD2 cells exhibited significantly less β-hexosaminidase release evoked by these substances compared with responses in control-siRNA-treated cells, while IgE-mediated release was comparable (Extended Data Fig. 10b). The residual β-hexosaminidase release observed in MRGPRX2-siRNA-treated cells is probably due to incomplete messenger RNA and/or protein knockdown.

Knowledge of the role of MRGPRX2 in drug-induced pseudo-allergies should expand further, for two reasons. First, ligand binding requirement studies should enable more specific screens for drugs that cross-activate MRGPRX2. Second, screening orally administered drugs may uncover more MRGPRX2 ligands, since common side effects of orally administered drugs include gastrointestinal problems and headache, both of which may have a mast cell component.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 7 March; accepted 28 October 2014.
Published online 17 December 2014.

1. Metcalfe, D. D., Baram, D. & Mekori, Y. A. Mast cells. *Physiol. Rev.* **77,** 1033–1079 (1997).
2. Galli, S. J., Nakae, S. & Tsai, M. Mast cells in the development of adaptive immune responses. *Nature Immunol.* **6,** 135–142 (2005).
3. Lagunoff, D., Martin, T. W. & Read, G. Agents that release histamine from mast cells. *Annu. Rev. Pharmacol. Toxicol.* **23,** 331–351 (1983).
4. Halpern, B. N. & Wood, D. R. The action of promethazine (phenergan) in protecting mice against death due to histamine. *Br. J. Pharmacol. Chemother.* **5,** 510–516 (1950).
5. Taneike, T., Miyazaki, H., Oikawa, S. & Ohga, A. Compound 48/80 elicits cholinergic contraction through histamine release in the chick oesophagus. *Gen. Pharmacol.* **19,** 689–695 (1988).
6. Ferry, X., Brehin, S., Kamel, R. & Landry, Y. G protein-dependent activation of mast cell by peptides and basic secretagogues. *Peptides* **23,** 1507–1515 (2002).
7. Purcell, W. M., Doyle, K. M., Westgate, C. & Atterwill, C. K. Characterisation of a functional polyamine site on rat mast cells: association with a NMDA receptor macrocomplex. *J. Neuroimmunol.* **65,** 49–53 (1996).
8. Tatemoto, K. *et al.* Immunoglobulin E-independent activation of mast cell is mediated by Mrg receptors. *Biochem. Biophys. Res. Commun.* **349,** 1322–1328 (2006).
9. Sick, E., Niederhoffer, N., Takeda, K., Landry, Y. & Gies, J. P. Activation of CD47 receptors causes histamine secretion from mast cells. *Cell. Mol. Life Sci.* **66,** 1271–1282 (2009).
10. Robas, N., Mead, E. & Fidock, M. MrgX2 is a high potency cortistatin receptor expressed in dorsal root ganglion. *J. Biol. Chem.* **278,** 44400–44404 (2003).
11. Subramanian, H., Gupta, K., Guo, Q., Price, R. & Ali, H. Mas-related gene X2 (MrgX2) is a novel G protein-coupled receptor for the antimicrobial peptide LL-37 in human mast cells: resistance to receptor phosphorylation, desensitization, and internalization. *J. Biol. Chem.* **286,** 44739–44749 (2011).
12. Kashem, S. W. *et al.* G protein coupled receptor specificity for C3a and compound 48/80-induced degranulation in human mast cells: roles of Mas-related genes *MrgX1* and *MrgX2*. *Eur. J. Pharmacol.* **668,** 299–304 (2011).
13. Subramanian, H. *et al.* β-Defensins activate human mast cells via Mas-related gene X2. *J. Immunol.* **191,** 345–352 (2013).
14. Kamohara, M. *et al.* Identification of MrgX2 as a human G-protein-coupled receptor for proadrenomedullin N-terminal peptides. *Biochem. Biophys. Res. Commun.* **330,** 1146–1152 (2005).
15. Liu, Q. *et al.* Sensory neuron-specific GPCR Mrgprs are itch receptors mediating chloroquine-induced pruritus. *Cell* **139,** 1353–1365 (2009).
16. Mousli, M., Hugli, T. E., Landry, Y. & Bronner, C. Peptidergic pathway in human skin and rat peritoneal mast cell activation. *Immunopharmacology* **27,** 1–11 (1994).
17. Pundir, P. & Kulka, M. The role of G protein-coupled receptors in mast cell activation by antimicrobial peptides: is there a connection? *Immunol. Cell Biol.* **88,** 632–640 (2010).
18. Hausmann, O., Schnyder, B. & Pichler, W. J. Etiology and pathogenesis of adverse drug reactions. *Chem. Immunol. Allergy* **97,** 32–46 (2012).
19. Lumry, W. R. *et al.* Randomized placebo-controlled trial of the bradykinin B_2 receptor antagonist icatibant for the treatment of acute attacks of hereditary angioedema: the FAST-3 trial. *Ann. Allergy Asthma Immunol.* **107,** 529–537 (2011).
20. Nel, L. & Eren, E. Peri-operative anaphylaxis. *Br. J. Clin. Pharmacol.* **71,** 647–658 (2011).
21. Read, G. W. Compound 48–80. Structure-activity relations and poly-THIQ, a new, more potent analog. *J. Med. Chem.* **16,** 1292–1295 (1973).
22. Mertes, P. M., Alla, F., Trechot, P., Auroy, Y. & Jougla, E. Anaphylaxis during anesthesia in France: an 8-year national survey. *J. Allergy Clin. Immunol.* **128,** 366–373 (2011).

23. Koppert, W. *et al.* Different patterns of mast cell activation by muscle relaxants in human skin. *Anesthesiology* **95,** 659–667 (2001).
24. Kelesidis, T., Fleisher, J. & Tsiodras, S. Anaphylactoid reaction considered ciprofloxacin related: a case report and literature review. *Clin. Ther.* **32,** 515–526 (2010).
25. Blanca-Lopez, N. *et al.* Hypersensitivity reactions to fluoroquinolones: analysis of the factors involved. *Clin. Exp. Allergy* **43,** 560–567 (2013).
26. Mori, K., Maru, C. & Takasuna, K. Characterization of histamine release induced by fluoroquinolone antibacterial agents *in-vivo* and *in-vitro*. *J. Pharm. Pharmacol.* **52,** 577–584 (2000).
27. Mori, K., Maru, C., Takasuna, K. & Furuhama, K. Mechanism of histamine release induced by levofloxacin, a fluoroquinolone antibacterial agent. *Eur. J. Pharmacol.* **394,** 51–55 (2000).
28. Doyle, E., Trosien, J. & Metz, M. Protocols for the induction and evaluation of systemic anaphylaxis in mice. *Methods Mol. Biol.* **1032,** 133–138 (2013).
29. Bergman, R. K. & Munoz, J. Efficacy of β-adrenergic blocking agents in inducing histamine sensitivity in mice. *Nature* **217,** 1173–1174 (1968).
30. Matsumura, Y., Tan, E. M. & Vaughan, J. H. Hypersensitivity to histamine and systemic anaphylaxis in mice with pharmacologic β adrenergic blockade: protection by nucleotides. *J. Allergy Clin. Immunol.* **58,** 387–394 (1976).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank C. Hawkins and H. Wellington for their assistance in generating MrgprB2-null mice and MrgprB2-Cre BAC transgenic mice. We thank B. Xiao for making MrgprB2-Cre BAC mice and Y. Geng for mouse genotyping. This work was supported by National Institutes of Health grants (R01NS054791 and R01GM087369 to X.D.). X.D. is an Early Career Scientist of the Howard Hughes Medical Institute.

Author Contributions B.D.M. conceived the project, designed and performed all experiments except where noted, and wrote the paper. P.P. performed all LAD2 cell work with supervision from M.K. S.M. performed tracheal contraction and tissue histamine release experiments. L.H. assisted with BAC purification and staining techniques. B.J.U. supervised S.M. and contributed to experimental design. X.D. supervised the project and wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to X.D. (xdong2@jhmi.edu).

METHODS

Animal models. All experiments were performed in accordance with a protocol approved by the Animal Care and Use Committee at the Johns Hopkins University School of Medicine. All experiments involving equal treatments in wild-type and mutant samples and animals were conducted by experimenters blind to conditions.

Analysis. Group data were expressed as mean ± s.e.m. Two-tailed unpaired Student's *t*-test was used to determine significance in statistical comparisons, and differences were considered significant at $P < 0.05$. Statistical power analysis was used to justify the sample size. We assumed the data were normally distributed since the most outcome values were symmetrically distributed around the mean value within each group. The variance is similar between groups as determined by the *F* test. Mast cells deemed to be damaged, either by visible lack of fibronectin adherence or by abnormally high resting calcium levels, were excluded from analysis. Otherwise, no samples or animals subjected to successful procedures and/or treatments were excluded from the analysis. No randomization was used for animal studies since it is not applicable for the studies.

Peptides and drugs. Compound 48/80, vespid mastoparan, rocuronium, tubocurarine, ciprofloxacin, levofloxacin, moxifloxacin and ofloxacin were from Sigma. Cortistatin was from Tocris Biosciences. PAMP(9–20) was custom synthesized and purified to ≥98% by Genscript. Leuprolide was from Genscript. Substance P, kallidin, mastoparan, cetrorelix, octreotide, sermorelin (growth hormone releasing factor 1–29) and icatibant (HOE-140) were from Anaspec. Atracurium and mivacurium were from Santa Cruz Biotechnology. Recombinant human insulin was from Roche. Goat anti-mouse IgE (Ab9162) was from Abcam.

Drug preparation and storage. Atracurium, mivacurium, tubocurarine and all fluoroquinolone solutions were prepared on the day of the experiment because the potencies of the first three were found to be susceptible to oxidation and/or freeze–thaw effects, while the solubility of the fluoroquinolones was best when prepared fresh. Propranolol was also prepared fresh on the day of the experiment to minimize the chances of a loss in potency. All fluoroquinolones except levofloxacin were dissolved into CIB adjusted to pH 3.5. All other drugs were prepared as 100×–1,000× aliquots and stored at −80 °C before thawing at 4 °C and diluting into calcium imaging buffer or saline.

Mrgpr RT–PCR screen. RNA was purified from 4×10^4 mouse peritoneal mast cells with a Qiagen RNEasy Micro column, according to the manufacturer's suggestions. RNA was treated for 20 min with DNase I (New England BioLabs) and re-purified on another RNEasy Micro column. Eight nanograms of RNA was used to generate first-strand complementary DNA using a SuperScript III kit (Invitrogen) according to the manufacturer's instructions, using oligo dT primers and scaling the recommended 10 µl reaction up to 60 µl. The negative control reaction was the same except that SuperScript III reverse transcriptase was replaced by water. Twenty-five microlitre PCR reactions were run with 12.5 µl RedTaq ReadyMix (Sigma), 0.5 µl dimethylsulphoxide (DMSO), 0.25 µl each of 50 µM gene-specific forward and reverse primers, 10 µl water and 2 µl mixture from the cDNA or negative control synthesis reactions. All reactions used a 4 min initial step at 95 °C, 30 s annealing at specific temperatures (described later), 40 s extension at 72 °C, and 25 s at 95 °C (with the last three steps repeated 39 times), and a final 4 min step at 72 °C. Low stringency PCR was set to 60 °C annealing; otherwise, annealing temperatures were: 62 °C for *Mrgpra1*, *Mrgpra10*, *Mrgprb2* and *Mrgprb6*; 64 °C for *Mrgpra2*, *Mrgpra3*, *Mrgpra4*, *Mrgpra6*, *Mrgpra16*, *Mrgpra18* and *Mrgprb11*; 65 °C for *Mrgpra9*, *Mrgpra19*, *Mrgprb1*, *Mrgprb3*, *Mrgprb5* and *Mrgprb8*; 66 °C for *Mrgpra12* and *Mrgprb10*; 63 °C for *Mrgprb4*; 61 °C for *Mrgpra14*; and 65.5 °C for *Mrgprc11*. Primers were as follows. *Mrgpra1*, forward, ATCCAGCAAGAGGAATGGGG, reverse, TGTGACCTAGGAGGAAGAAGAAG; *Mrgpra2*, forward, CCTCCTACACAAGCCAGCAA, reverse, AAGCACAAGTGAAAGATGATGCT; *Mrgpra3*, forward, GCTACATCCAGCAAGAGGAATG, reverse, GCAAAAATTCCTTTGGGTAGGGT; *Mrgpra4*, forward, CCTGTGTGCTGTGATCTGGT, reverse, TCACGGTTAATCCAGGGCAC; *Mrgpra6*, forward, CATTTTCCTCCCCCAACAGT, reverse, ATGCCTGAATGAGCCCACAA; *Mrgpra9*, forward, CAGTGATCTACATCCAGCAAAAGG, reverse, GCGTGGAAGCTATGATGCGA; *Mrgpra10*, forward, CAGTGGTCCACCATCTCCAA, reverse, ACAGGCAAGAGAGTCATGGTT; *Mrgpra12*, forward, TCAGGGATCGGGTGAAGCAC, reverse, GAGCATTTGAAGGTGTTGTTGGA; *Mrgpra14*, forward, GGTTGCCCCTGTGTTTCTTC, reverse, TATTGCCAGTCAGTAAGCTGAG; *Mrgpra16*, forward, GCCCTCTGGTTCCCATTACT, reverse, GTTTTTGGACCACTGAGGCATT); *Mrgpra18*, forward, TGCTCTGGTTTTCTCCTTTGC, reverse, TGAGGCATGTCAAGTCAGTCA; *Mrgpra19*, forward, CAGGACCCAGATCACGACAC, reverse, TCCTGGGCTTCCGATTTCAC; *Mrgprb1*, forward, ATTAGCCTTCATCAGGCACCA, reverse, CCAGCCCAACTAAGGCAATG; *Mrgprb2*, forward, GTCACAGACCAGTTTAACACTTCC, reverse, CAGCCATAGCCAGGTTGAGAA; *Mrgprb3*, forward, ACCTGGCTGTGGCTGATTTT, reverse, GCTGAACCCACAGAGAACCA; *Mrgprb4*, forward, TCTGGCTGGTGCTGATTTCTT, reverse, ACCACGAGGCTCAACAATAGA; *Mrgprb5*, forward, CTGTGGTTCCTTCTGTGTCCA, reverse, TTTCCAGTTCCCCAGACCTTT; *Mrgprb6*, forward, TCTGTCTACATCCTCAACCTGG, reverse, ATTATCTCATGAGGAAGGCTCAA; *Mrgprb8*, forward, AGAGAATGCAAAGCATGCGA, reverse, GAGGAAGTTTGCCCCAGACA; *Mrgprb10*, forward, CACTGGTCACATTGCCAACC, reverse, GGGGATGGAATCAATGTCCAAGA; *Mrgprb11*, forward, ACCTTCTTGCTATTTTTCCCTCCA, reverse, AGGATGAGACTGGACCCACA; *Mrgprc11*, forward, CAGCACAAGTCAGCTCCTCAA, reverse, ATGCCCATGAGAAAGGACAGAACC.

Expression constructs. *Mrgpr* genes were cloned and inserted into the pcDNA3.1 mammalian expression plasmid using standard techniques. All mouse genes had a Kozak sequence at their amino terminus and also encoded a carboxy-terminal Flag tag separated from the genes by the amino acid linker DIIL.

cDNA constructs. First-strand cDNA was prepared as described for RT–PCR screens, and amplification was performed using the Q5 HotStart High Fidelity Master Mix (New England Biolabs). At least five different clones each prepared from wild-type and mutant mice were sequenced to verify the presence of the deletion in the mutant and the absence of any other mutation from wild type or mutant.

Calcium imaging in HEK293 cells. In initial screens, HEK293 cells (not tested for mycoplasma but rapidly dividing) were transiently transfected with gene constructs including a C-terminal Flag tag, and plated on 100 µg ml^{-1} poly-D-lysine-coated glass coverslips 6 h after transfection. Twenty-four hours later, cells were loaded with AM esters of the calcium indicators Fura-2 or Fluo-4 (Molecular Probes) along with 0.02% Pluronic F-127 (Molecular Probes) for 45 min at 37 °C. Fura-2-loaded cells were imaged during 340 and 380 nm excitation, and Fluo-4 loaded cells were imaged during 488 nm excitation. Later experiments used cell lines stably expressing receptors along with transient or stable expression of the promiscuous G protein Gα15. Cells were imaged in calcium imaging buffer (CIB; NaCl 125 mM, KCl 3 mM, $CaCl_2$ 2.5 mM, $MgCl_2$ 0.6 mM, HEPES 10 mM, glucose 20 mM, $NaHCO_3$ 1.2 mM, sucrose 20 mM, brought to pH 7.4 with NaOH). Unless otherwise specified, drugs were perfused into the chamber for 45 to 60 s and responses were monitored at 5-s intervals for an additional 60–90 s.

EC_{50} determination. HEK293 cells stably expressing Gα15 and either Mrgprb2 or MRGPRX2 were plated at 4×10^4 cells per well in 96-well plates and incubated overnight. The next day, media was removed and replaced with imaging solution from the FLIPR Calcium 5 assay kit (Molecular Devices), diluted according to manufacturer's suggestions in Hank's balanced salt solution (HBSS) with 20 mM HEPES, pH 7.4. Cells were incubated in 100 µl imaging solution at 37 °C for 60 min, and allowed to recover for 15 min at room temperature before imaging in a Flexstation 3 (Molecular Devices). Wells were imaged according to manufacturer's specifications for 120 s, with 50 µl of test substances at three times the concentration added 30 s after imaging began. Responses were determined by subtracting the minimum signal from the maximum signal. Substances were tested in duplicate wells, the signals were averaged, and EC_{50} values were determined for each trial by normalizing to the peak response to the substance in that trial. All drugs were dissolved in HBSS plus HEPES solution, with the following exceptions due to solubility issues: cetrorelix acetate was dissolved in saline containing 2.5 mM $CaCl_2$ and 0.6 mM $MgCl_2$, and fluoroquinolones except ofloxacin were dissolved in the same solution except that the pH was adjusted with HCl to 3.5; ofloxacin required 100 µg ml^{-1} of lactic acid for full solubility. We also noticed that peptides sometimes lost potency after a freeze–thaw cycle, so most peptides were prepared directly from lyophilized stock.

Peritoneal mast cell purification and imaging. Adult male and female mice 2–5 months of age were killed through CO_2 inhalation. A total of 12 ml of ice-cold mast cell dissociation media (MCDM; HBSS with 3% fetal bovine serum and 10 mM HEPES, pH 7.2) were used to make two sequential peritoneal lavages, which were combined and cells were spun down at 200*g*. The pellet from each mouse was resuspended in 2 ml MCDM, layered over 4 ml of an isotonic 70% Percoll suspension (2.8 ml Percoll, 320 µl 10× HBSS, 40 µl 1 M HEPES, 830 µl MCDM), and spun down for 20 min, 500*g*, 4 °C. Mast cells were recovered in the pellet. Purity was >95%, as assayed by avidin staining and by morphology. Mast cells were resuspended at 5×10^5–1×10^6 cells ml^{-1} in DMEM with 10% fetal bovine serum and 25 ng ml^{-1} recombinant mouse stem cell factor (Sigma), and plated onto glass coverslips coated with 30 µg ml^{-1} fibronectin (Sigma). For counting, instead of plating, suspended mast cells were diluted 1/10 and affixed to slides by spinning at 1,000 r.p.m. for 5 min at 4 °C on a CytoSpin (Thermo Scientific).

For imaging, after 2 h of incubation at 37 °C, 5% CO_2, mast cells were loaded with Fluo-4 along with 0.02% Pluronic F-127 for 30 min at room temperature, washed three times in CIB and used immediately for imaging. Cells were used within 2 h of loading. Cells were identified as responding if the $[Ca^{2+}]_i$ rose by at least 50% for at least 10 s, which clearly distinguishes a ligand-induced response from random flickering events. Average traces were calculated by taking the average response from each cell in a mouse, and averaging those.

BAC transgenic mice generation. We purchased the BAC clone RP23-65I23 from the Children's Hospital Oakland Research Institute. This clone contains the *Mrgprb2*

locus, ~60 kb of 5′ genomic sequence and over 100 kb of 3′ genomic sequence. Recombineering in bacteria was used to introduce eGFP-Cre and a polyA signal immediately after the *Mrgprb2* start codon[31]. The BAC was linearized with NotI (New England Biolabs) and injected into pronuclei from single-cell-fertilized C57BL/6 eggs. Eggs were implanted into pseudopregnant females. Three BAC mouse lines were established. Although mice were already in a C57BL/6 background, they were crossed for at least four generations to wild-type and tdTomato reporter mice in the C57BL/6 background before use in experiments. BAC mice were mated to *ROSA26*^{tdTomato} mice purchased from Jackson Laboratories for imaging studies. Experiments for Fig. 1 used mice homozygous for *ROSA26*^{tdTomato} because the tdTomato signal was often heterogeneous and weak in heterozygous mice. Genotyping reactions for BAC mice were run at 61 °C annealing, and primers were: forward, TATATCATGGCCGACAAGCA; reverse, CAGACCGCGCGCCTGAAGA. Both primers are in the *eGFP-Cre* reading frame but the entire gene and correct placement in the *Mrgprb2* locus was verified by previous sequencing.

Mrgprb2 mutant mice generation. mRNAs encoding zinc finger nucleases targeting *Mrgprb2* were purchased from Sigma. The binding sites were GTTCCTGGGCATCCG and TGCACACGAATGCCTTCACTG, corresponding to bases 180–194 and 196–216, respectively, of the *Mrgprb2* open reading frame. mRNA was diluted to 2 ng ml^{-1} in 1 mm Tris-HCl buffer, pH 7.4, with 0.25 mm EDTA, and injected into the pronuclei of single-cell-fertilized eggs in the C57BL/6 strain. No overt signs of toxicity were observed. Embryos were implanted into pseudopregnant females. DNA flanking the binding sites was amplified from founder mice and screened for mutations using the Cel-1 assay kit (Transgenomics), according to the manufacturer's suggestions. Three of the first 28 mice were identified and confirmed by DNA sequencing to carry small mutations, and no more screening was performed. In addition to the 4 bp mutation used in this study, a mouse carrying a 1 bp deletion and another with a 2 bp deletion were identified.

Wild-type and Mrgprb2^{MUT} mouse genotyping. Primers used for wild-type mice were GGTTCCTGGGCATCCGTAT and GGTTCCTGGGCATCCGTAT, and reactions were run at an annealing temperature of 62.8 °C. Primers for Mrgprb2^{MUT} mice were GTTCCTGGGCATCCGCAC and CTTCCGCCTGAACCTTCGGT, and reactions were run at 64.0 °C annealing temperature.

Avidin labelling of tissue. Adult male and female mice up to 8 months of age were anaesthetized with pentobarbital and perfused with 20 ml 0.1 M PBS (pH 7.4, 4 °C) followed by 25 ml of fixative (4% formaldehyde (vol/vol), 4 °C). Heart, trachea and skin sections were dissected from the perfused mice. Tissues were post-fixed in fixative at 4 °C overnight. When skin sections were the only tissues needed, they were dissected and placed in fixative directly after asphyxiation of mice by CO_2 inhalation, eliminating the perfusion step. Tissues were cryoprotected in 20% sucrose (wt/vol) for more than 24 h and were sectioned (20 μm width) with a cryostat. The sections on slides were dried at 37 °C for 30 min, and fixed with 4% paraformaldehyde at 21–23 °C for 10 min. The slides were pre-incubated in blocking solution (10% normal goat serum (vol/vol), 0.2% Triton X-100 (vol/vol) in PBS, pH 7.4) for 1 or 2 h at 21–23 °C, then incubated with 1/500 FITC-avidin (Sigma) or rhodamine-avidin (Vector Labs) for 45 min. Sections were washed three times with water or PBS and a drop of Fluoromount G (SouthernBiotech) was added before coverslips were placed on top. Heart mast cells were examined near cavities because the density was much higher than elsewhere in the tissue; avidin-positive, tdTomato-negative cells were observed embedded in muscle tissue in very low numbers, but their identity was unclear.

For avidin labelling of peritoneal mast cells, cells were plated as described earlier, fixed with 4% paraformaldehyde at 21–23 °C for 10 min, incubated with 1/1,000 avidin in PBS for 30 min at 21–23 °C, and washed with PBS before immediate imaging.

Stomach section immunocytochemistry. Adult male and female mice up to 8 months of age were anaesthetized with pentobarbital and perfused with 20 ml 0.1 M PBS (pH 7.4, 4 °C) followed by 25 ml of fixative (4% formaldehyde (vol/vol), 4 °C). Stomach sections were removed, washed thoroughly, postfixed in 4% formaldehyde for 2 h, and prepared for sectioning by incubation in a 30% sucrose solution for 48 h. Tissue samples were mounted in cryoembedding media and frozen, and 14 μm sections were made using a crytostat and then fixed onto slides. Slides were washed with a 0.2% Triton X-100 PBS solution, incubated for 1 h in a 10% normal goat serum solution, and then incubated overnight at 4 °C with a 1:20 dilution of rat monoclonal anti-mouse MCPT1 (monoclonal antibody RF6.1, eBiosciences) in a 0.2% Triton/1% normal goat serum solution. Slides were washed with the 0.2% Triton solution and incubated for 2 h at room temperature in Triton solution with a 1:500 dilution of a goat anti-rat IgG Alexa Fluor 488 conjugated antibody (Life Technologies). Slides were washed in PBS before coverslips were added with an anti-fade solution for imaging.

Peripheral white blood cell preparation. Blood was collected from Mrgprb2-tdTomato mice via cardiac punctures with a syringe containing PBS with 30 units ml^{-1} heparin and 5 mM EDTA, diluted 1:1 with the same solution, and allowed to cool to room temperature before layering over 6 ml of a Histopaque-1119 solution in a 15 ml conical tube. Tubes were centrifuged at 700*g* for 30 min, and white blood cells were collected at the interface between the PBS and Histopaque solutions. Cells were washed with PBS and spun down at 500*g* for 10 min a total of three times. Cells were spun onto poly-lysine-coated slides in a Cytospin 4 (Thermo Scientific) at 600 r.p.m. for 3–5 min, dried overnight on a 37 °C heating block, and incubated for 2 min with Hoechst 33342 diluted to 0.5 μg ml^{-1} in PBS before coverslip mounting with an anti-fade solution. In parallel, we also stained cells in suspension with Hoechst 33342, spun the cells down, and mixed the resuspended cells directly in a PBS/anti-fade solution before placing directly onto slides and mounting coverslips on the suspension. No tdTomato-positive cells were seen in any preparation using either method.

Tissue histamine release studies. Whole tracheae or segments of skin isolated from the abdominal aspect of shaved male and female mice up to 6 months of age (4–8 mg wet weight) were dissected and cleaned of connective tissue. After a 60 min incubation period in oxygenated Krebs' bicarbonate buffer solution (37 °C), the tissue was treated with either vehicle or compound 48/80 for 30 min. The supernatant solution was saved for histamine analysis. The tissue was then subjected to 8% percholoric acid in a 37 °C water bath for 15 min to obtain total histamine content. Histamine was assayed by the automated fluorometric technique previously described[32].

Tracheal contractions. Tracheal contractions were carried out as previously described[33]. For allergen (ovalbumin) responses, mice were actively sensitized by injecting 0.2 ml of an ovalbumin solution (3.75 μg ml^{-1}) mixed with $Al(OH)_3$ three times at an interval of 2 days. Experiments were conducted on male and female animals 8–12 weeks of age beginning 2 weeks after the first injection. Trachea were cleaned of connective tissue and tracheal rings (whole or laterally divided in half), were suspended between two tungsten stirrups in 10 ml organ chambers filled with Krebs' buffer that was warmed to 37 °C and bubbled with 95% O_2–5% CO_2 to maintain a pH of 7.4. One stirrup was connected to a strain gauge (model FT03; Grass Instruments), and tension was recorded on a Grass Model 7 polygraph (Grass Instruments). Preparations were stretched to a resting tension of 0.2*g*, and washed with fresh Krebs' buffer at 15-min intervals during a 60-min equilibration period. After equilibration, trachea were challenged with either ovalbumin (10 μg ml^{-1}) or compound 48/80. At the end of each experiment, all trachea were maximally contracted with carbachol (1 μM). All results are expressed as a percentage of maximum contraction.

Hindpaw swelling and extravasation. Adult male mice up to 8 months of age were anaesthetized with an intraperitoneal (i.p.) injection of 50 mg kg^{-1} pentobarbital (Sigma). Fifteen minutes after induction of anaesthesia, mice were injected intravenously (i.v.) with 50 μl of 12.5 mg ml^{-1} Evans blue (Sigma) in saline. Five minutes later, 5 μl of the test substance (or 7 μl of anti-IgE) was administered by intraplantar injection in one paw and saline was administered in the other paw. Paw thickness was measured by callipers immediately after injection. Fifteen minutes later (30 min after anti-IgE), paw thickness was measured again and mice were killed by decapitation. Paw tissue was collected, dried for 24 h at 50 °C, and weighed. Evans blue was extracted by a 24 h incubation in formamide at 50 °C, and the OD was read at 620 nm using a spectrophotometer. For studies using ketotifen, mice were injected i.p. with 25 μl of a 10 mg ml^{-1} solution of ketotifen at the same time as pentobarbital.

Systemic anaphylaxis assay. To minimize stress, animals were transported to the procedure area the day before injections. Adult male and female mice up to 8 months of age (25 to 35 g) were given an i.p. injection of 80 μg propranolol in saline (2 mg ml^{-1}) immediately after removal from their cages, and then placed back in their cages for 30 min before intravenous injections. The intravenous injections were performed on one mouse at a time. For each injection, a mouse was placed in a transport box and brought to a room with no other mice, to minimize stress from vocalizations during injection. The mouse was then placed in a restrainer, and the injection was performed within 4 min of restraint because we observed that longer restraint times affected body core temperature independent from the injection. Tail veins were dilated by repeated wiping of tail with a tissue soaked in 100% ethanol, followed by injection of ciprofloxacin in a 0.25 ml Hamilton syringe fit with a 30.5-gauge needle (BD Biosciences). The injection was determined to be successful only when all of the criteria were met: blood appeared in the syringe after needle insertion, all tail veins were visible after injection, and the mouse bled slightly from the injection site after needle withdrawal. The injection site was swabbed until blood stopped flowing, the mouse was placed in a separate cage from its housing cage, one mouse per cage, and returned to the room it was brought from. At least one wild-type and one mutant mouse were used for each experimental session. Body core temperature was measured with a rectal thermometer.

Mouse peritoneal mast cell histamine release assay. Mast cells were purified as described earlier and allowed to recover for 2 h in DMEM with 10% FBS and 25 ng ml^{-1} mouse stem cell factor in a 37 °C incubator with 5% CO_2. Cells were

then spun down, resuspended in CIB, counted, and plated at 300 cells per well in 75 µl CIB in 96-well plates coated with 20 µg ml^{-1} fibronectin (Sigma). They were allowed to adhere to the substrate for 45 min at 37 °C in atmospheric conditions (that is, CO_2 levels were not adjusted) before assay. For the assays, cells were removed to room temperature and 75 µl of 2× concentrations of tested substances (all in CIB except for ciprofloxacin, which was in saline with 2.5 mM $CaCl_2$ and 0.6 mM $MgCl_2$, pH 3.5) were added. After 5 min, 40 µl of supernatant was aspirated, diluted with 40 µl CIB and frozen at −80 °C until histamine levels were determined. Anti-IgE treatment was similar, except that cells were incubated for 30 min at 37 °C after anti-IgE was added before aspiration of supernatant. Histamine content was determined by using an HTRF histamine assay kit (Cisbio Assays) according to the manufacturer's instructions.

Human mast cell culture. LAD2 (Laboratory of Allergic Diseases 2) human mast cells were cultured in StemPro-34 SFM medium (Life Technologies) supplemented with 2 mM L-glutamine, 100 U ml^{-1} penicillin, 50 µg ml^{-1} streptomycin and 100 ng ml^{-1} recombinant human stem cell factor (Peprotech). The cell suspensions were seeded at a density of 0.1×10^6 cells ml^{-1} and maintained at 37 °C and 5% CO_2, and periodically tested for the expression of CD117 and FcεRI by flow cytometry. Cell culture medium was hemi-depleted every week with fresh medium.

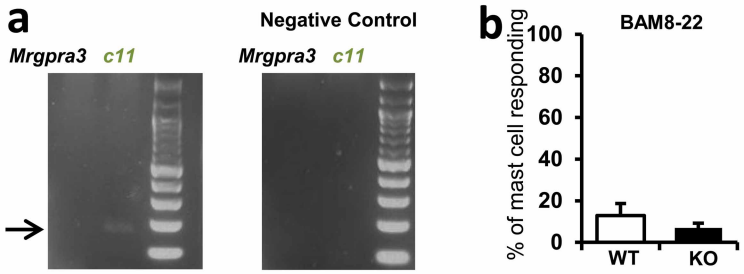
LAD2 degranulation assay. LAD2 cells were sensitized for 20 h with 0.5 µg ml^{-1} biotin-conjugated human IgE (Abbiotec). Cells were washed, resuspended in HEPES buffer (10 mM HEPES, 137 mM NaCl, 2.7 mM KCl, 0.38 mM $Na_2HPO_4.7H_2O$, 5.6 mM glucose, 1.8 mM $CaCl_2.H_2O$, 1.3 mM $MgSO_4.7H_2O$, 0.4% BSA, pH 7.4) at 0.025×10^6 per well, and then stimulated with 0.1 µg ml^{-1} streptavidin (Life Technologies) or other agonists at the indicated concentrations for 30 min at 37 °C/5% CO_2. The β-hexosaminidase released into the supernatants and in cell lysates was quantified by hydrolysis of p-nitrophenyl *N*-acetyl-β-D-glucosamide (Sigma-Aldrich) in 0.1 M sodium citrate buffer (pH 4.5) for 90 min at 37 °C. The percentage of β-hexosaminidase release was calculated as a per cent of total content. Agonists tested were compound 48/80, mastoparan, icatibant, atracurium bessylate and ciprofloxacin hydrochloride.

Enzyme immunoassay and ELISA. LAD2 cells were washed with medium, suspended at 0.25×10^6 cells per well, and incubated with compound 48/80, mastoparan, icatibant, atracurium or ciprofloxacin at the indicated concentrations for 3–24 h at 37 °C/5% CO_2. Cell-free supernatants were harvested and analysed for PGD_2 release by an enzyme immunoassay (Cayman chemical), while TNF content was quantified using an ELISA kit (eBioscience) according to the manufacturer's instruction. The minimum detection limits were 55 pg ml^{-1} for PGD_2 and 5.5 pg ml^{-1} for TNF.

Measurement of histamine release from LAD2 cells. LAD2 cells were washed, suspended in BSA-free HEPES buffer at 0.1×10^6 per well, and incubated with compound 48/80, mastoparan, icatibant, atracurium or ciprofloxacin at the indicated concentrations for 30 min at 37 °C/5% CO_2. A histamine (Sigma-Aldrich) stock solution of 100 µg ml^{-1} was prepared and stored at −20 °C. The working standards of 4,000 ng ml^{-1} to 7.8 ng ml^{-1} were freshly prepared using twofold serial dilution. *O*-phthalaldehyde (OPT; Sigma-Aldrich) was dissolved in acetone-free methanol (10 mg ml^{-1}) and kept in the dark at 4 °C. Histamine standards and cell-free supernatants (60 µl) were transferred to a flat-bottom 96-black-well microplate and mixed with 12 µl 1 M NaOH and 3 µl OPT. After 4 min at room temperature, 6 µl 3 M HCl was added to stop the histamine-OPT reaction. Fluorescence intensity was measured using a 355 nm excitation filter and a 460 nm emission filter.

siRNA transfection of LAD2 cells. Expression of *MRGPRX2* was downregulated with ON-TARGET plus SMARTpool siRNA against *MRGPRX2* and control siRNA from Dharmacon. LAD2 cells were washed with medium, suspended at 0.5×10^6 cells per well, and transfected with 100 nm *MRGPRX2* siRNA and control siRNA in antibiotic-free StemPro medium using Lipofectamine 3000 (Life Technologies) according to the manufacturer's instruction at 37 °C/5% CO_2. At 48 h, knockdown was confirmed by RT–PCR, and the cells were used for degranulation assays.

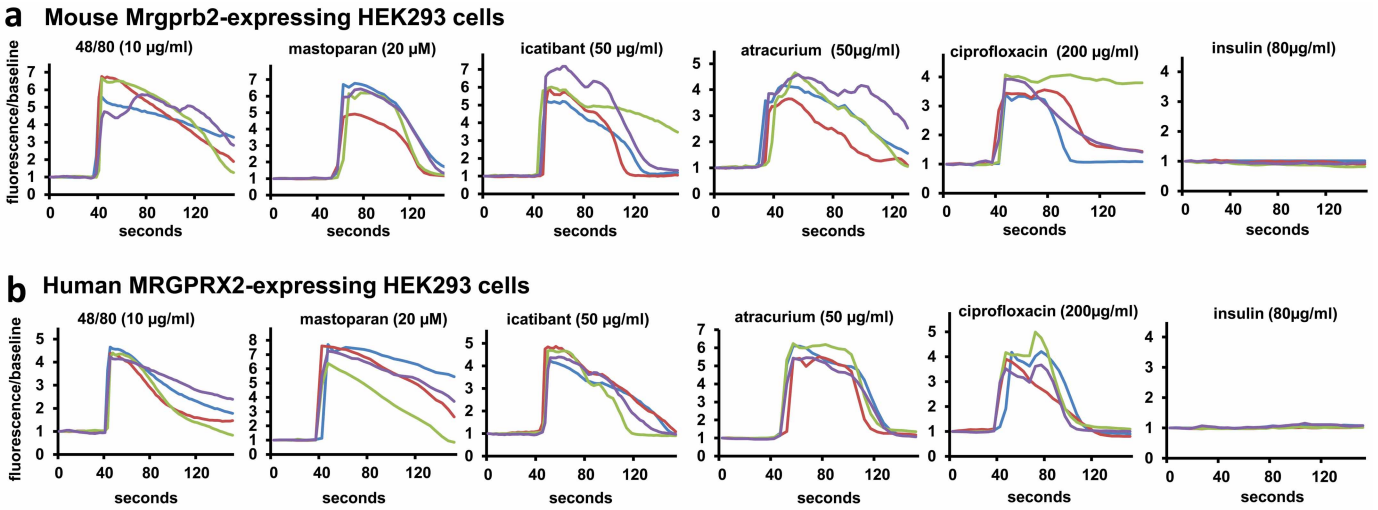
31. Han, L. *et al.* A subpopulation of nociceptors specifically linked to itch. *Nature Neurosci.* **16,** 174–182 (2013).
32. Siraganian, R. P. An automated continuous-flow system for the extraction and fluorometric analysis of histamine. *Anal. Biochem.* **57,** 383–394 (1974).
33. Weigand, L. A., Myers, A. C., Meeker, S. & Undem, B. J. Mast cell-cholinergic nerve interaction in mouse airways. *J. Physiol. (Lond.)* **587,** 3355–3362 (2009).
34. Maurer, M. *et al.* Mast cells promote homeostasis by limiting endothelin-1-induced toxicity. *Nature* **432,** 512–516 (2004).



c

	X2		b2	b1	b10	b11
PAMP (50 μM)	✓		✓	×	×	×
48/80 (20 μg/ml)	✓		✓	×	×	×
Cort. (20 μM)	✓		✓	×	×	×
Sub. P (100 μM)	✓		✓	×	×	×
CQ (100 μM)	×		×	×	×	×

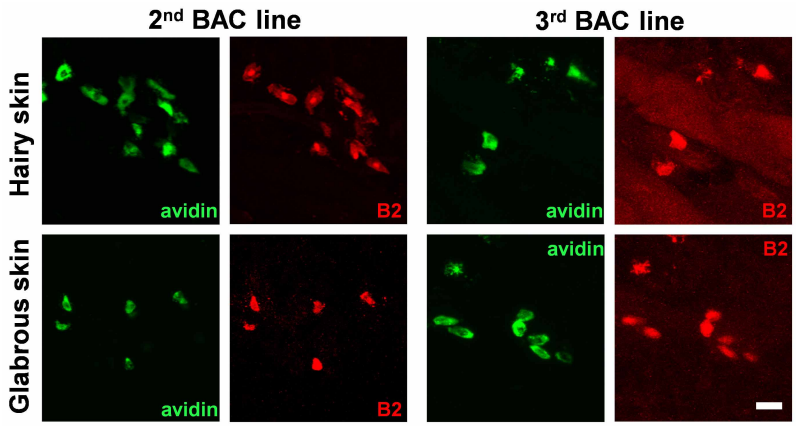
Extended Data Figure 1 | MRGPRX1 orthologues are not expressed at relevant levels in mast cells under naive conditions. **a**, Results from a low-stringency RT–PCR screen (see Methods) in peritoneal mast cells for expression of the *MRGPRX1* orthologues *Mrgpra3* and *Mrgprc11*. Arrow points to expected band sizes. **b**, Percentages of peritoneal mast cells responding to the MRGPRX1 and Mrgprc11 agonist bovine adrenal medulla derived peptide, fragment 8–22 (BAM8–22, 500 nM). Activation was assayed by measuring rises in intracellular calcium, using imaging of the Fluo-4 dye. Differences are not significant ($P = 0.39$). $n = 3$ mice from each genotype. Group data are expressed as mean ± s.e.m. Two-tailed unpaired Student's *t*-test was used to determine significance in statistical comparisons. WT, wild type; KO, knockout. **c**, Chart summarizing responses to MRGPRX2 ligands and the MRGPRX1 ligand chloroquine (CQ) by HEK293 cells transiently transfected with plasmids driving expression of MRGPRX2, Mrgprb2 and other mouse Mrgpr proteins (that is, Mrgprb1, Mrgprb10 and Mrgprb11) most closely related to Mrgprb2. Positive and negative responses are indicated with ticks and crosses, respectively. Responses were considered positive if at least half of the transfected cells showed a 50% increase in $[Ca^{2+}]_i$. No cells transfected with Mrgprb1, Mrgprb10 and Mrgprb11 responded to any listed drug.



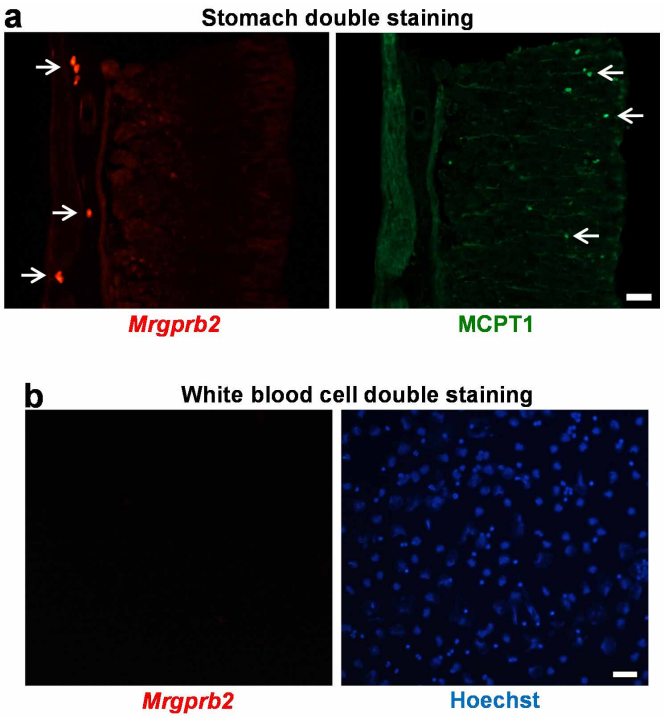
c

Substance	Mrgprb2 EC_{50}	MRGPRX2 EC_{50}
Compound 48/80	3.7 ± 0.5 μg/ml	470.1 ± 139.6 ng/ml
Substance P	54.3 ± 4.9 μM	152.3 ± 48.0 nM
Cortistatin-14	21.3 ± 0.9 μM	106.7 ± 39.3 nM
PAMP (9-20)	12.4 ± 1.6 μM	166.0 ± 35.7 nM
Mastoparan	24.0 ± 3.6 μM	3.9 ± 0.7 μM
Icatibant	32.5 ± 2.0 μg/ml	15.8 ± 2.7 μg/ml
Cetrorelix	23.4 ± 1.4 μg/ml	221.7 ± 63.1 ng/ml
Sermorelin	29.1 ± 1.2 μg/ml	4.5 ± 0.9 μg/ml
Octreotide	10.0 ± 1.1 μg/ml	6.6 ± 0.7 μg/ml
Leuprolide	152.0 ± 7.1 μg/ml	9.1 ± 0.7 μg/ml
Atracurium	44.8 ± 1.4 μg/ml	28.6 ± 2.4 μg/ml
Rocuronium	22.2 ± 3.3 μg/ml	261.3 ± 14.4 μg/ml
Ciprofloxacin	126.5 ± 5.1 μg/ml	6.8 ± 0.5 μg/ml
Moxifloxacin	14.1 ± 2.1 μg/ml	9.9 ± 0.6 μg/ml
Levofloxacin	807.6 ± 47.1 μg/ml	22.7 ± 0.4 μg/ml
Ofloxacin	225.0 ± 25.4 μg/ml	30.1 ± 1.5 μg/ml

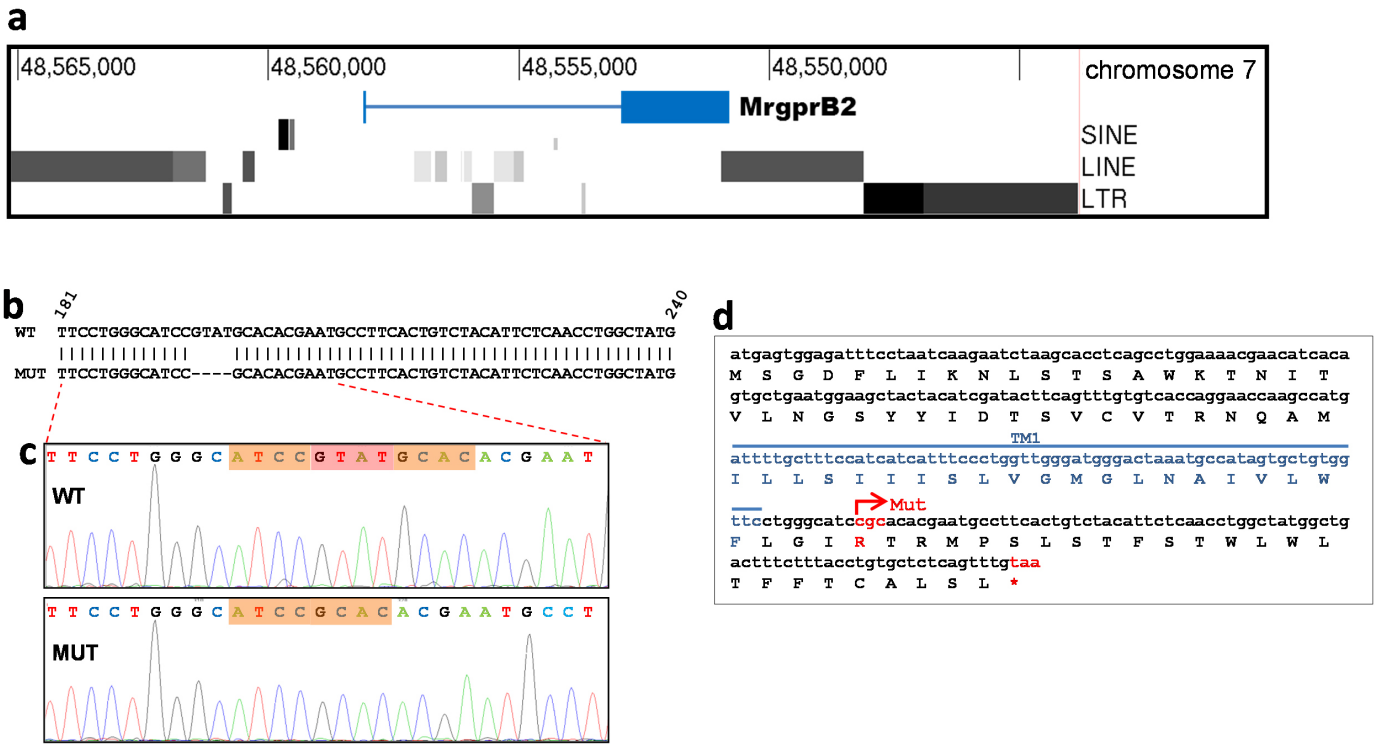
Extended Data Figure 2 | Basic secretagogues and drugs that induce pseudo-allergic reactions activate mouse Mrgprb2 and human MRGPRX2 expressed in HEK293 cells. **a**, **b**, Example traces showing changes in $[Ca^{2+}]_i$, as measured by Fluo-4 imaging, from HEK293 cells expressing Mrgprb2 and Gα15 (**a**) or MRGPRX2 and Gα15 (**b**). Substances were perfused from the 30 to 90 s time period, except for ciprofloxacin, which was perfused between the 30 and 60 s time periods to minimize exposure to the low pH solutions it was dissolved in. Insulin was used as a negative control. **c**, Table of half-maximum effective concentration (EC_{50}) values of basic secretagogues and drugs associated with pseudo-allergic reactions to activate Mrgprb2- and MRGPRX2-expressing HEK293 cells. The EC_{50} values were determined from dose–response studies, which were repeated three times. Data are expressed as mean ± s.e.m.



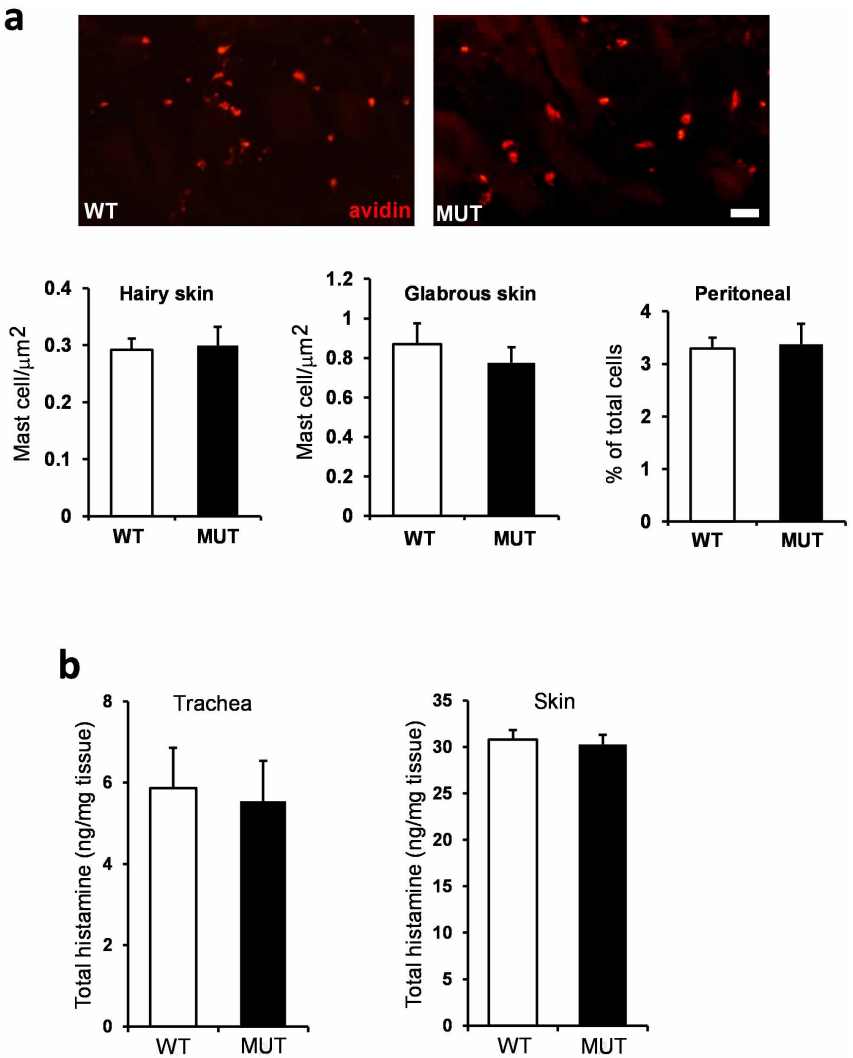
Extended Data Figure 3 | Multiple lines of BAC transgenic mice confirm mast-cell-specific MrgprB2 expression. Representative confocal images from two other BAC transgenic mouse lines. BAC mice expressing eGFP-Cre in the *Mrgprb2* open reading frame were mated to tdTomato reporter mice and tdTomato (red) expression was compared to avidin staining (green), a marker for mast cells. Scale bar, 20 μm.



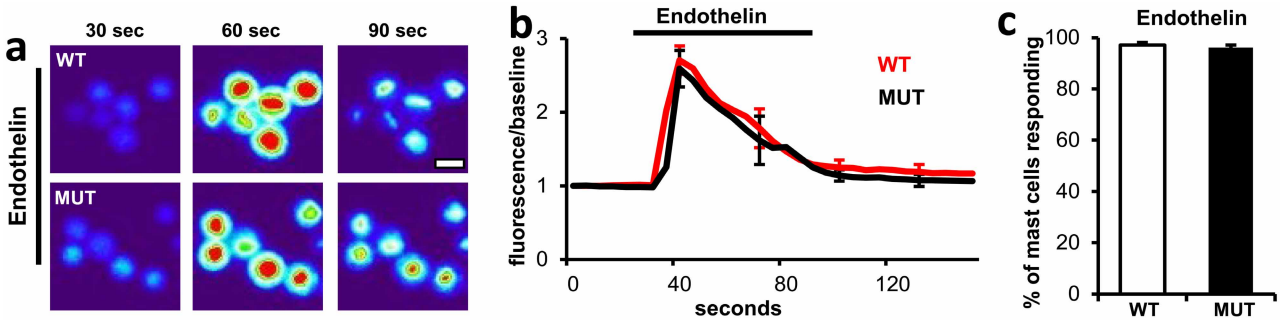
Extended Data Figure 4 | Mrgprb2 is not expressed in mucosal mast cells or peripheral white blood cells. **a**, Representative images of a stomach section from an Mrgprb2-tdTomato mouse stained with an anti-MCPT1 (β-chymase) antibody to label mucosal mast cells. White arrows indicate positive cells. No cells were double-labelled (296 Mcpt1-labelled cells and 275 tdTomato-positive cells counted, $n = 3$ mice). Scale bar, 40 μm. **b**, Representative images of a Cytospin preparation of peripheral white blood cells from an Mrgprb2-tdTomato mouse doubly labelled with tdTomato for Mrgprb2-expressing cells (red; left image) and Hoechst 33342 nuclear staining (blue; right image). No peripheral white blood cell expressed tdTomato ($n = 3$ mice; >4,000 cells examined). Scale bar, 40 μm.



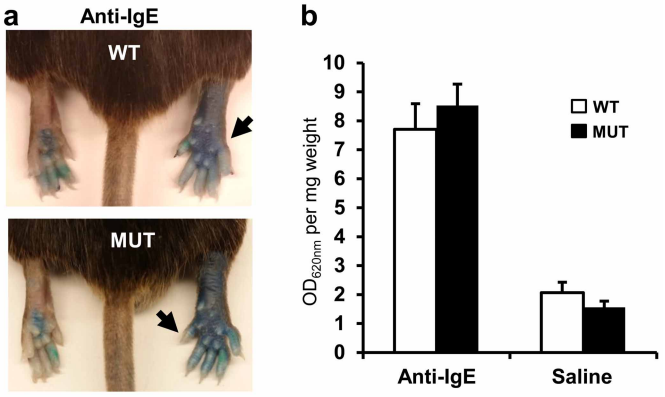
Extended Data Figure 5 | Mrgprb2^{MUT} mice are functional knockouts. **a**, Illustration of the genomic region in and around the *Mrgprb2* locus. Note that repetitive sequences including long interspersed elements (LINEs), short interspersed elements (SINEs), and long tandem repeats (LTRs) begin immediately after the 3′ side of the *Mrgprb2* gene, and in addition are present within 2.5 kb of the 5′ side. A BLASTN search in March 2014 using the 500 bases adjacent to the 3′ end of *Mrgprb2* as a query turned up more than 269,000 hits in the mouse genome. **b**, Comparison of the wild-type (WT) and mutant (MUT) genomic sequences shows the location of the 4 bp deletion in the mutant. Numbers correspond to the *Mrgprb2* open reading frame. **c**, Sequencing result from wild-type and mutant complementary DNA sampled from mice born 18 months after the mutant line was established. The bases missing in the mutant are highlighted in red. **d**, Amino acid translation of the *Mrgprb2*^{MUT} open reading frame reveals that the deletion creates a frameshift mutation and an early termination codon (marked with an asterisk) shortly after the first transmembrane region. Mut, site of the frameshift deletion; TM1, transmembrane region 1.



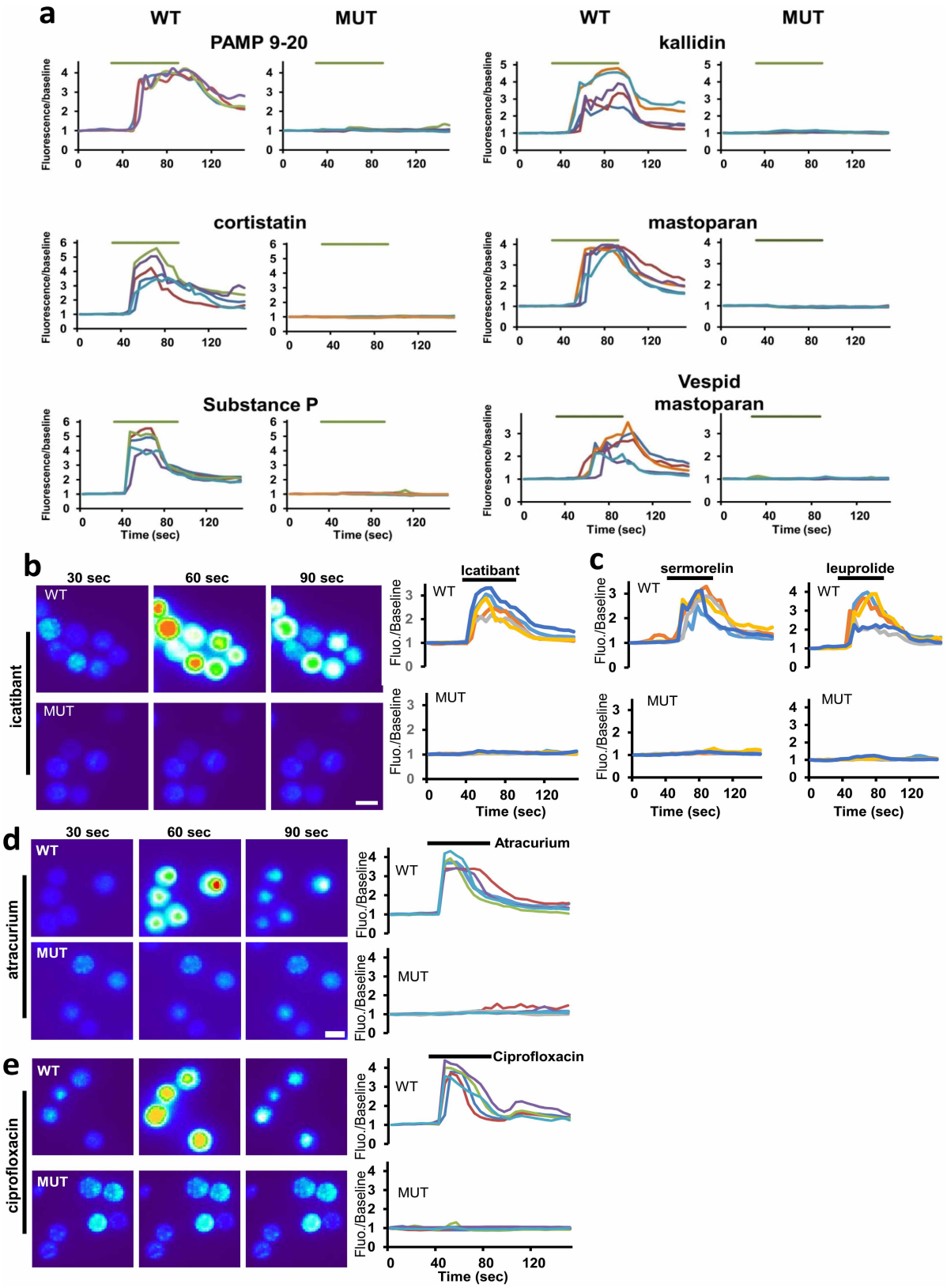
Extended Data Figure 6 | The mast cell numbers and the histamine content of tracheal and skin tissue was not different between wild-type and Mrgprb2^{MUT} animals. **a**, Top, representative pictures of avidin staining in wild-type (WT) and Mrgprb2^{MUT} (MUT) mice. Scale bar, 40 μm. Bottom, quantification of mast cell numbers in various tissues. Differences are not significant, using a two-tailed unpaired Student's *t*-test ($n = 3$ mice for each genotype; over 3,000 μm^2 and 1,000 μm^2 counted for each genotype for hairy and glabrous skin, respectively; over 10,000 peritoneal cells counted). **b**, The tracheal histamine content averaged 5.9 ± 0.9 and 5.5 ± 1.6 ng mg^{-1} ($n = 5$ for each genotype), respectively; the skin histamine content averaged 30.8 ± 3.2 and 30.2 ± 4.0 ng mg^{-1} ($n = 8$ for each genotype), respectively. Differences were not significant. Group data are expressed as mean ± s.e.m. Two-tailed unpaired Student's *t*-test was used to determine significance in statistical comparisons.



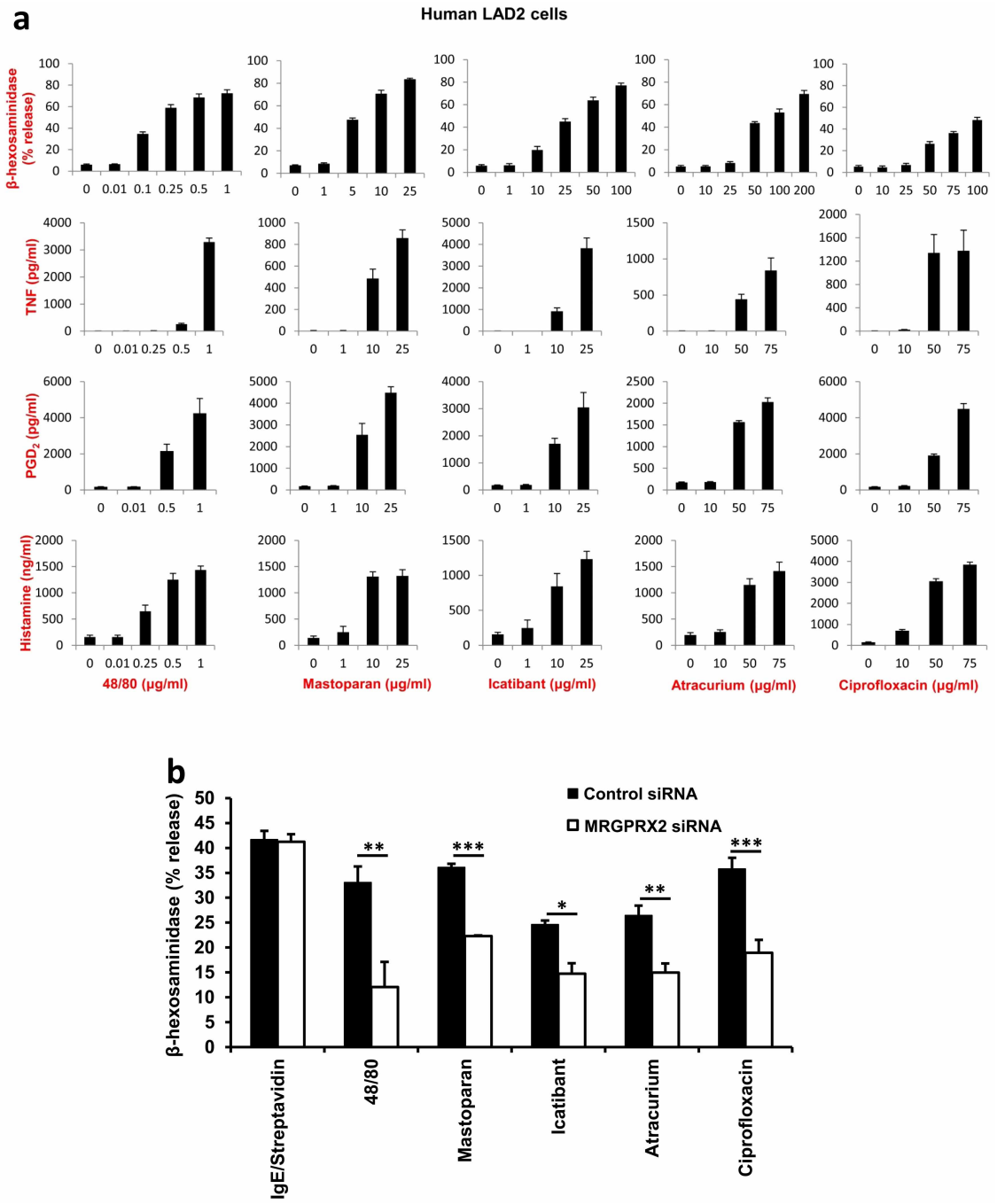
Extended Data Figure 7 | Endothelin acting through the ETA GPCR[34] induced comparable activation in Mrgprb2^{MUT} and wild-type mast cells. **a**, Representative heat map images of mouse peritoneal mast cells showing changes in $[Ca^{2+}]$i, as assayed by Fluo-4 imaging, induced by bath application of endothelin (1 μM). Scale bar, 10 μm. **b**, Averages of $[Ca^{2+}]$i imaging traces for wild-type (WT) (red line) and Mrgprb2^{MUT} (MUT) (black line). The $[Ca^{2+}]$i traces are similar between wild-type and mutant groups. Traces were averaged as described for Fig. 2a. **c**, Quantification of percentage of responding cells. Group data are expressed as mean ± s.e.m. Two-tailed unpaired Student's *t*-test was used to determine significance in statistical comparisons ($n = 3$ for each genotype; over 180 cells counted for each genotype). Endothelin-induced responses were not significantly different.



Extended Data Figure 8 | IgE-mediated inflammation does not differ between wild-type and MrgprB2^{MUT} mice. **a**, Representative images of Evans blue stained extravasation 15 min after intraplantar injection of anti-IgE antibody (right, arrow, 100 μg ml^{-1}, 7 μl in saline) or saline (left). **b**, Quantification of Evans blue leakage into the paw after 15 min ($n = 6$ for wild type (WT), $n = 7$ for MrgprB2^{MUT} (MUT)). Differences after anti-IgE antibody ($P = 0.49$) and saline ($P = 0.23$) injection are not significant. Group data are expressed as mean ± s.e.m. Two-tailed unpaired Student's *t*-test was used to determine significance in statistical comparisons.



Extended Data Figure 9 | Mrgprb2^{MUT} mast cells are unresponsive to basic secretagogues and various therapeutic drugs. **a**, Example traces showing changes in $[Ca^{2+}]_i$, as measured by Fluo-4 imaging, from wild-type (WT) and Mrgprb2^{MUT} (MUT) peritoneal mast cells induced by the basic secretagogues from Fig. 2e. Each trace is a response from a unique cell. **b**, Representative Fluo-4 images (left) and fluorescence traces (right) from wild-type (top) and Mrgprb2^{MUT} (bottom) cultured peritoneal mast cells during application of icatibant (50 μg ml^{-1}). **c**, Example traces showing changes in $[Ca^{2+}]_i$, as measured by Fluo-4 imaging, from wild-type and Mrgprb2^{MUT} peritoneal mast cells induced by selected FDA-approved cationic peptidergic drugs. Each trace is a response from a unique cell. **d**, Representative Fluo-4 images (left) and fluorescence traces (right) from wild-type (top) and Mrgprb2^{MUT} (bottom) cultured peritoneal mast cells during application of atracurium (50 μg ml^{-1}). **e**, Representative Fluo-4 images (left) and fluorescence traces (right) from wild-type (top) and Mrgprb2^{MUT} (bottom) cultured peritoneal mast cells during application of ciprofloxacin (200 μg ml^{-1}).



Extended Data Figure 10 | Human mast cells are activated by basic secretagogues and drugs associated with pseudo-allergic reactions in an MRGPRX2-dependent manner. a, Human LAD2 mast cells were treated with different concentrations of compound 48/80, mastoparan, icatibant, atracurium and ciprofloxacin. The activation of mast cells in response to these substances was characterized by the release of β-hexosaminidase, TNF, PGD_2 and histamine. In addition, 0.1 μg ml^{-1} streptavidin stimulation of biotin-conjugated human IgE-sensitized LAD2 cells caused a robust release of β-hexosaminidase (71.3 ± 1.8% release), compared with untreated cells (4.1 ± 0.3% release). Group data are expressed as mean ± s.e.m. **b**, Knockdown of human *MRGPRX2* significantly reduced mast cell activation evoked by basic secretagogues and drugs associated with pseudo-allergic reactions, but not by IgE. Human LAD2 mast cells were first transfected with siRNA against *MRGPRX2* or control siRNA. Two days after the transfection, the cells were treated with compound 48/80 (0.1 μg ml^{-1}), mastoparan (5 μg ml^{-1}), icatibant (10 μg ml^{-1}), atracurium (25 μg ml^{-1}) and ciprofloxacin (75 μg ml^{-1}). The activation of mast cells in response to these substances characterized by the release of β-hexosaminidase was significantly reduced in MRGPRX2-siRNA-treated cells, compared to release in the control group. IgE-mediated mast cell degranulation was unaffected by MRGPRX2 siRNA knockdown. Group data are expressed as mean ± s.e.m. Two-tailed unpaired Student's *t*-test was used to determine significance in statistical comparisons, and differences were considered significant at $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.005$ (the experiments were repeated three times).

LETTER

doi:10.1038/nature14115

Group 2 innate lymphoid cells promote beiging of white adipose tissue and limit obesity

Jonathan R. Brestoff[1,2], Brian S. Kim[2]†, Steven A. Saenz[2]†, Rachel R. Stine[3], Laurel A. Monticelli[1,2], Gregory F. Sonnenberg[1], Joseph J. Thome[4,5], Donna L. Farber[4,5,6], Kabirullah Lutfy[7], Patrick Seale[3] & David Artis[1,2]

Obesity is an increasingly prevalent disease regulated by genetic and environmental factors. Emerging studies indicate that immune cells, including monocytes, granulocytes and lymphocytes, regulate metabolic homeostasis and are dysregulated in obesity[1,2]. Group 2 innate lymphoid cells (ILC2s) can regulate adaptive immunity[3,4] and eosinophil and alternatively activated macrophage responses[5], and were recently identified in murine white adipose tissue (WAT)[5] where they may act to limit the development of obesity[6]. However, ILC2s have not been identified in human adipose tissue, and the mechanisms by which ILC2s regulate metabolic homeostasis remain unknown. Here we identify ILC2s in human WAT and demonstrate that decreased ILC2 responses in WAT are a conserved characteristic of obesity in humans and mice. Interleukin (IL)-33 was found to be critical for the maintenance of ILC2s in WAT and in limiting adiposity in mice by increasing caloric expenditure. This was associated with recruitment of uncoupling protein 1 (UCP1)$^+$ beige adipocytes in WAT, a process known as beiging or browning that regulates caloric expenditure[7–9]. IL-33-induced beiging was dependent on ILC2s, and IL-33 treatment or transfer of IL-33-elicited ILC2s was sufficient to drive beiging independently of the adaptive immune system, eosinophils or IL-4 receptor signalling. We found that ILC2s produce methionine-enkephalin peptides that can act directly on adipocytes to upregulate *Ucp1* expression *in vitro* and that promote beiging *in vivo*. Collectively, these studies indicate that, in addition to responding to infection or tissue damage, ILC2s can regulate adipose function and metabolic homeostasis in part via production of enkephalin peptides that elicit beiging.

Group 2 innate lymphoid cells (ILC2s) respond to the cytokine interleukin (IL)-33 (refs 3,10,11), and both IL-33 and ILC2s have been implicated in the regulation of metabolic homeostasis in mice[5,6,12]. To address whether ILCs are present in human white adipose tissue (WAT) or dysregulated in obese patients, we obtained abdominal subcutaneous WAT from non-obese human donors and identified a lineage (Lin)-negative cell population that expresses CD25 (IL-2Rα) and CD127 (IL-7Rα) (Fig. 1a, Extended Data Fig. 1a). This cell population expressed GATA binding protein 3 (GATA-3) and the IL-33 receptor (IL-33R) (Fig. 1b), consistent with ILC2s in other human tissues[13,14]. A Lin$^-$ CD25$^+$ CD127$^+$ cell population that expresses GATA-3 and IL-33R was also identified in epididymal (E)-WAT of mice (Fig. 1c, d). These cells were developmentally dependent on inhibitor of DNA binding 2 (Id2), transcription factor 7 (TCF-7) and the common gamma chain (γ_c) and produced the effector cytokines IL-5 and IL-13 (Extended Data Fig. 1b–e), similar to murine ILC2s as described previously[3,5,10,11,14,15].

We compared ILC2 frequencies in abdominal subcutaneous WAT from non-obese versus obese donors (Extended Data Table 1). WAT from obese donors exhibited decreased frequencies of ILC2s compared to non-obese controls (Fig. 1e, f). The obese group was enriched in older females compared to the non-obese group, but age and sex did not explain the difference in ILC2 frequencies between obese and non-obese donors (Extended Data Fig. 1f, g). To test whether ILC2s in WAT are also dysregulated in murine obesity, mice were fed a control diet or high-fat diet (HFD). HFD-induced obese mice exhibited decreased frequencies and numbers of ILC2s in E-WAT compared to wild-type mice fed a control diet (Fig. 1g, h). Together, these data suggest that decreased ILC2 populations in WAT is a conserved characteristic of obesity in mice and humans.

We employed IL-33-deficient mice to test whether endogenous IL-33 regulates ILC2 responses and the development of obesity. *Il33*$^{-/-}$ mice exhibited decreased basal frequencies and numbers of ILC2s in E-WAT and inguinal (i)WAT compared to *Il33*$^{+/+}$ controls (Fig. 2a–c, Extended Data Fig. 2a), and expression of IL-5 and IL-13 by WAT ILC2s was decreased in *Il33*$^{-/-}$ mice compared to controls (Extended Data Fig. 2b). Notably, when fed a normal diet, mice lacking IL-33 gained more weight, accumulated more E-WAT and iWAT and had increased adipocyte size and whole-body adiposity compared to controls (Fig. 2d–f, Extended Data Fig. 2c). In addition, *Il33*$^{-/-}$ mice exhibited dysregulated glucose homeostasis as evidenced by fasting euglycaemic hyperinsulinaemia, increased HOMA-IR index (homeostatic model assessment of insulin resistance) values and impaired glucose and insulin tolerance (Extended Data Fig. 2d–h). Together, these results indicate that endogenous IL-33 is required to maintain normal ILC2 responses in WAT and to limit the development of spontaneous obesity.

In contrast, wild-type mice treated with recombinant murine (rm)IL-33 exhibited increased accumulation of ILC2s in E-WAT and iWAT (Fig. 2g–i). Although body weight did not differ between groups (Fig. 2j), mice treated with rmIL-33 had decreased adiposity and increased lean mass compared to controls (Fig. 2k). Remarkably, HFD-fed mice treated with rmIL-33 displayed increased E-WAT ILC2 numbers in association with decreased body weight and fat mass and improved glucose homeostasis compared to HFD-fed mice treated with PBS (Extended Data Fig. 3a–f). These beneficial metabolic effects are consistent with studies showing a protective role for IL-33 in obesity[12] and may be related to obesity-associated pathologies such as atherosclerosis that are limited by IL-33[16].

To examine the mechanisms by which IL-33 regulates adiposity we assessed energy homeostasis in control and rmIL-33-treated mice. Treatment of mice with rmIL-33 for 7 days resulted in increased caloric expenditure compared to controls (Fig. 2l). Food intake was unchanged following chronic rmIL-33 treatment (Fig. 2m), and the absence of hyperphagia in the setting of increased caloric expenditure seemed to be related to decreased activity (Fig. 2n, Extended Data Fig. 4a). However, rmIL-33 did not appear to have direct suppressive effects on food intake or activity levels (Extended Data Fig. 4b–d). These data suggest

[1]Jill Roberts Institute for Research in IBD, Joan and Sanford I. Weill Department of Medicine, Department of Microbiology and Immunology, Weill Cornell Medical College, Cornell University, New York, New York 10021, USA. [2]Department of Microbiology and Institute for Immunology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. [3]Institute for Diabetes, Obesity and Metabolism, Department of Cell and Developmental Biology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. [4]Columbia Center for Translational Immunology, Columbia University Medical Center, New York, New York 10032, USA. [5]Department of Microbiology and Immunology, Columbia University Medical Center, New York, New York 10032, USA. [6]Department of Surgery, Columbia University Medical Center, New York, New York 10032, USA. [7]Department of Pharmaceutical Sciences, College of Pharmacy, Western University of Health Sciences, Pomona, California 91766, USA. †Present addresses: Division of Dermatology, Department of Medicine, Washington University School of Medicine, St Louis, Missouri 63110, USA (B.S.K.); Immunology Research, Biogen Idec, Inc., Cambridge, Massachusetts 02142, USA (S.A.S.).

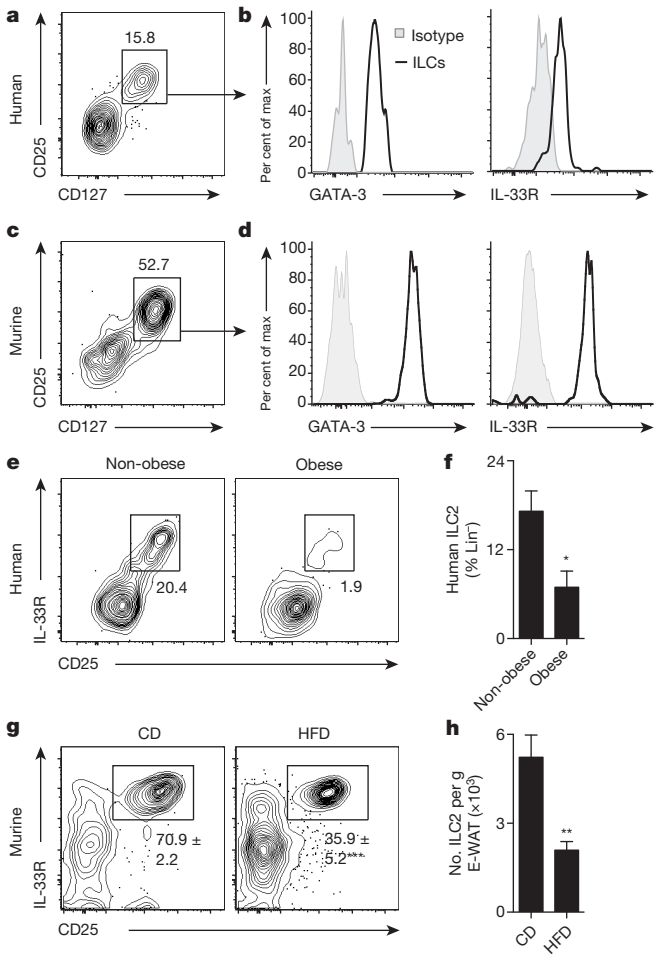


Figure 1 | Human and murine white adipose tissue contains group 2 innate lymphoid cells that are dysregulated in obesity. **a**, Identification of lineage (Lin)-negative CD25$^+$ CD127$^+$ innate lymphoid cells (ILCs) in human abdominal subcutaneous white adipose tissue (WAT) of a lean donor. Pre-gated on live CD45$^+$ Lin$^-$ cells that lack CD3, CD5, TCRαβ, CD19, CD56, CD11c, CD11b, CD16, and FcεRIα. **b**, Histograms of GATA-3 and IL-33R expression by human WAT ILCs (line). Shaded histogram, isotype control. **c**, Identification of Lin$^-$ CD25$^+$ CD127$^+$ ILCs in murine epididymal (E)-WAT. Pre-gated on live CD45$^+$ Lin$^-$ cells that lack CD3, CD5, CD19, NK1.1, CD11c, CD11b and FcεRIα. **d**, Histograms of GATA-3 and IL-33R expression by murine E-WAT ILCs (line). Shaded histogram, isotype control. **e**, Representative plots and **f**, frequencies of human WAT ILC2s from donors stratified into non-obese (body mass index (BMI) $< 30.0\,kg\,m^{-2}$, $n = 7$) and obese (BMI $\geq 30.0\,kg\,m^{-2}$, $n = 7$) groups. **g**, Representative plots and frequencies of murine E-WAT ILC2s from mice fed a control diet (CD, 10% kcal fat, $n = 5$) or high-fat diet (HFD, 45% kcal fat, $n = 4$) for 12 weeks. **h**, Numbers of murine ILC2s per gram of E-WAT in mice fed a CD ($n = 8$) or HFD ($n = 6$) for 12 weeks. Student's t-test, $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$. Data are shown as mean ± standard error and are representative of 2–3 independent experiments. Sample sizes are biological replicates.

that increased caloric expenditure following 7 days of rmIL-33 treatment could not be explained by the thermic effect of food or physical activity levels, but was regulated by other physiologic processes.

An emerging cell type that is critical for regulating caloric expenditure is the beige adipocyte (also known as brite, brown-like or inducible brown adipocyte)[7,9,17,18]. These specialized adipocytes produce heat by uncoupling energy substrate oxidation from ATP synthesis[7,17,18], a thermogenic process that expends calories and is dependent on uncoupling protein 1 (UCP1)[8,17]. Previous work has linked brown and beige adipocyte function to the prevention of weight gain in mice and humans[9,19–21]. To test whether IL-33 regulates beiging, we examined WAT morphology of *Il33*$^{+/+}$ versus *Il33*$^{-/-}$ mice. iWAT from *Il33*$^{+/+}$ mice exhibited unilocular white adipocytes with interspersed paucilocular beige adipocytes that have multiple small lipid droplets and increased UCP1$^+$ cytoplasm (Fig. 3a). In contrast, iWAT from *Il33*$^{-/-}$ mice had few beige adipocytes (Fig. 3b) and increased white adipocyte size compared to controls (Fig. 3a, b, Extended Data Fig. 2c). Expression of *Ucp1* was also lower in iWAT of *Il33*$^{-/-}$ mice compared to controls (Fig. 3c), suggesting that IL-33 may be a critical regulator of beiging. Consistent with this, mice treated with rmIL-33 exhibited increased UCP1$^+$ beige adipocytes and elevated expression of *Ucp1* messenger RNA in E-WAT and iWAT (Fig. 3d–f) compared to controls, indicating that IL-33 can promote beiging of WAT. Notably, the stimulatory effect of rmIL-33 treatment on UCP1 expression was restricted to WAT and was not observed in brown adipose tissue (BAT) (Extended Data Fig. 5a–e).

To test whether IL-33-elicited ILC2s promote beiging, congenic CD45.1$^+$ ILC2s from E-WAT of IL-33-treated donor mice were sort-purified and transferred into wild-type CD45.2$^+$ recipient mice (Extended

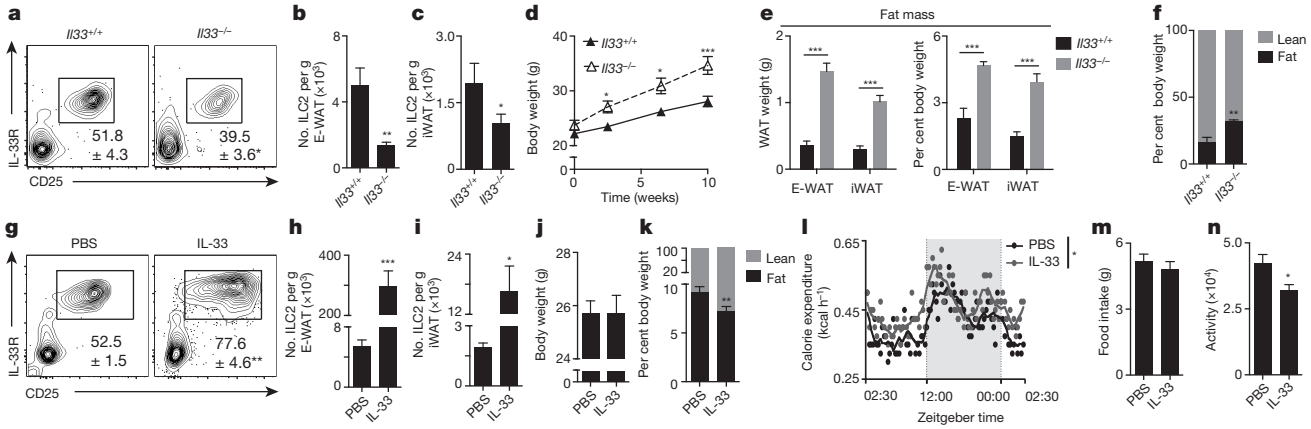


Figure 2 | IL-33 critically regulates ILC2 responses in white adipose tissue and limits adiposity. **a**–**f**, *Il33*$^{+/+}$ ($n = 6$) or *Il33*$^{-/-}$ ($n = 5$) mice were fed a control diet (10% kcal fat) for 12 weeks starting at 7 weeks of age. **a**, Frequencies and **b**, numbers of live CD45$^+$ Lin$^-$ CD25$^+$ IL-33R$^+$ ILC2s in epididymal (E)-WAT. Plots pre-gated on CD45$^+$ Lin$^-$ cells that lack CD3, CD5, CD19, NK1.1, CD11c, CD11b and FcεRIα. **c**, Numbers of ILC2s in inguinal (i)WAT. **d**, Body weight, first 10 weeks of feeding. **e**, Absolute and relative E-WAT and iWAT weights. **f**, Body composition. **g**–**n**, Wild-type mice were treated with phosphate buffered saline (PBS, $n = 10$) or recombinant murine IL-33 (12.5 μg per kg body weight per day, $n = 12$) by intraperitoneal injection for 7 days. **g**, Frequencies and **h**, numbers of ILC2s in E-WAT. **i**, Numbers of ILC2s in iWAT. **j**, Body weight and **k**, body composition. **l**, Caloric expenditure over a 24-h period, days 6-to-7 of treatment. Non-shaded area, lights on. Shaded area, lights off. **m**, Food intake and **n**, total activity (beam breaks) over the 24-h period in **l**. Student's t-test or ANOVA with repeated measures. $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$. Data are shown as mean ± standard error and are representative of 2 independent experiments. Sample sizes are biological replicates.

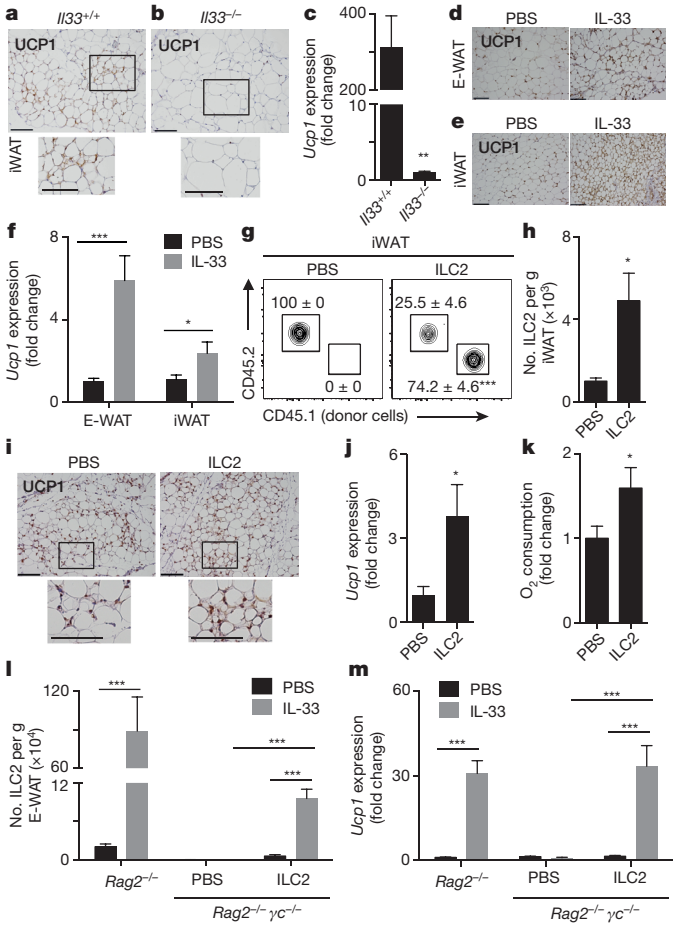


Figure 3 | IL-33 and ILC2s contribute to beiging of white adipose tissue. **a**–**c**, *Il33*$^{+/+}$ ($n = 6$) or *Il33*$^{-/-}$ ($n = 5$) mice were fed a low-fat diet (10% kcal fat) for 12 weeks starting at age 7 weeks. Uncoupling protein 1 (UCP1) immunohistochemistry (IHC) in iWAT from **a**, *Il33*$^{+/+}$ or **b**, *Il33*$^{-/-}$ mice. Scale bars, 100 μm. **c**, *Ucp1* transcript levels in iWAT. **d**–**f**, Wild-type mice were treated with PBS or recombinant murine IL-33 (12.5 μg per kg body weight per day) by intraperitoneal injection for 7 days. **d**, E-WAT and **e**, iWAT UCP1 IHC. Scale bars, 100 μm. **f**, *Ucp1* transcript levels in E-WAT and iWAT. **g**–**k**, Sort-purified CD45.1^{+} ILC2s ($\times 10^5$) from E-WAT of IL-33-treated mice were transferred into 12-week-old CD45.2^{+} wild-type recipients by subcutaneous and intraperitoneal injection daily for 4 days (PBS, $n = 8$; ILC2, $n = 8$ except panel **k**). **g**, Representative plots identifying donor and recipient ILC2s. Plots pre-gated on live CD45^{+} Lin^{-} CD25^{+} IL33R^{+} cells. Lineage cocktail: CD3, CD5, CD19, NK1.1, CD11c, CD11b and FcεRIα. **h**, Total numbers of ILC2s per gram iWAT. **i**, iWAT UCP1 IHC. Scale bars, 100 μm. **j**, *Ucp1* expression in iWAT. **k**, iWAT oxygen consumption. PBS, $n = 14$; ILC2, $n = 15$. **l**, **m**, Sort-purified congenic CD45.1^{+} ILC2s ($\times 10^5$) from E-WAT of IL-33-treated mice were transferred into *Rag2*$^{-/-}$ *γc*$^{-/-}$ mice once by intraperitoneal injection. ILC2-sufficient *Rag2*$^{-/-}$ mice, ILC2-deficient *Rag2*$^{-/-}$ *γc*$^{-/-}$ mice and ILC2-reconstituted *Rag2*$^{-/-}$ *γc*$^{-/-}$ mice were treated with PBS or recombinant murine IL-33 (12.5 μg per kg body weight per day) by intraperitoneal injection for 7 days ($n = 4$ mice per group). **l**, ILC2 numbers per gram E-WAT. **m**, *Ucp1* expression in E-WAT. Student's *t*-test or two-way ANOVA. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$. Data are shown as mean ± standard error and are representative of 2–4 independent experiments. Sample sizes are biological replicates.

Data Fig. 6a). CD45.1^{+} donor ILC2s could be identified in iWAT (Fig. 3g) and E-WAT (Extended Data Fig. 6b) of mice that received ILC2s but not in control mice that received PBS, and total ILC2 numbers were significantly increased in iWAT of mice receiving CD45.1^{+} ILC2s compared to controls (Fig. 3h). Transferred ILC2s could not be identified in BAT, mesenteric lymph nodes or lung (Extended Data Fig. 6b), indicating selective accumulation of WAT-derived ILC2s in WAT of recipient mice. Transfer of ILC2s was associated with increased UCP1^{+} beige adipocytes, augmented expression of *Ucp1* and elevated oxygen consumption in iWAT (Fig. 3i–k).

To test whether IL-33 promotes beiging of WAT in an ILC2-dependent manner, we treated ILC2-deficient *Rag2*$^{-/-}$ *γc*$^{-/-}$ mice with IL-33 in the presence or absence of adoptively transferred congenic ILC2s (Extended Data Fig. 6c). *Rag2*$^{-/-}$ *γc*$^{-/-}$ (*γc* is also known as *Il2rg*) mice supported accumulation and IL-33-induced population expansion of transferred E-WAT-derived ILC2s in host E-WAT (Fig. 3l, Extended Data Fig. 6d). IL-33 treatment increased expression of *Ucp1* in E-WAT of ILC2-sufficient *Rag2*$^{-/-}$ controls but not ILC2-deficient *Rag2*$^{-/-}$ *γc*$^{-/-}$ mice (Fig. 3m). Strikingly, rmIL-33-induced increases in expression of *Ucp1* and beiging were restored in ILC2-reconstituted *Rag2*$^{-/-}$ *γc*$^{-/-}$ mice (Fig. 3m, Extended Data Fig. 6e). Collectively, these results indicate that IL-33-induced beiging of WAT requires a γ_c-dependent cell population and that ILC2s are sufficient to rescue this defect, suggesting that IL-33-induced beiging is critically dependent on ILC2s.

ILC2s have been shown to promote the eosinophil/IL-4Rα/alternatively-activated macrophage (AAMac) pathway that can elicit beiging through IL-4Rα-dependent production of noradrenaline by AAMacs[5,22–24]. In addition, regulatory T (T_{reg}) cells in WAT are known to be critical for regulating glucose homeostasis in mice[25] and are increased following rmIL-33 treatment (Extended Data Fig. 3g, h). Therefore, we sought to test whether the IL-33/ILC2 pathway could promote beiging in the absence of eosinophils, IL-4Rα or the adaptive immune system. Remarkably, delivery of rmIL-33 to *DblGata1* (also known as *Gata1*^{tm6Sho}; eosinophil-deficient), *Il4ra*$^{-/-}$ or *Rag2*$^{-/-}$ mice elicited beiging of WAT (Fig. 3m, Extended Data Fig. 7a–f), and transfer of IL-33-elicited ILC2s to *DblGata1*, *Il4ra*$^{-/-}$ or *Rag1*$^{-/-}$ mice resulted in accumulation of ILC2s in iWAT and recruitment of UCP1^{+} beige adipocytes (Extended Data Fig. 7g–l). Therefore, although eosinophils, AAMacs and adaptive immune cells may contribute to optimal beiging under some physiologic settings, these data indicate that the IL33/ILC2 axis can promote beiging independently of the eosinophil/IL-4Rα/AAMac pathway and the adaptive immune system.

Obesity is associated with both decreased ILC2s (Fig. 1) and defective beige adipocytes[5,9,21]. To address whether ILC2s produce factors that could directly regulate beiging, we employed genome-wide transcriptional profiling of ILC2s versus group 3 ILCs (ILC3s) to compare gene expression enrichment scores of 69 genes previously linked to human obesity (Extended Data Table 2)[26,27]. This analysis identified one gene, proprotein convertase subtilisin/kexin type 1 (*Pcsk1*) (also known as prohormone convertase 1, PC1), to be significantly enriched in ILC2s but not ILC3s (Fig. 4a, $P < 0.01$). PCSK1 is an endopeptidase involved in processing some prohormones into active forms[28], and loss-of-function mutations in both mice and humans are associated with increased susceptibility to obesity and decreased caloric expenditure[29]. The most differentially expressed PCSK1 target in ILC2s versus ILC3s was proenkephalin A (*Penk*) (Fig. 4b), which encodes endogenous opioid-like peptides such as methionine-enkephalin (MetEnk). Production of MetEnk by ILC2s was confirmed by flow cytometric analysis of sort-purified ILC2s (Fig. 4c). Following IL-33 stimulation, ILC2 production of MetEnk peptides was increased (Fig. 4d). *In vivo* delivery of MetEnk peptides into wild-type mice elicited UCP1^{+} beige adipocytes, upregulated expression of *Ucp1* and increased oxygen consumption in iWAT (Fig. 4e–g), indicating the formation of functional beige fat. Consistent with this, MetEnk treatment decreased iWAT mass (Fig. 4h). These changes were not associated with increased expression of *Il4* or *Il13* (Fig. 4i) or altered eosinophil or AAMac numbers in iWAT (Fig. 4j).

Gene expression analyses in wild-type mice at steady state indicated that *Il33* and *Penk* expression levels were increased in iWAT compared to BAT (Fig. 4k). In addition, expression of the MetEnk receptor δ1 opioid receptor (*Oprd1*) was higher in iWAT compared to BAT, whereas expression of the other known MetEnk receptor Opioid growth factor receptor (*Ogfr*) was lower in iWAT compared to BAT (Fig. 4l), suggesting that there may be tissue-specific effects of MetEnk in iWAT compared to BAT. Consistent with this, MetEnk stimulation induced

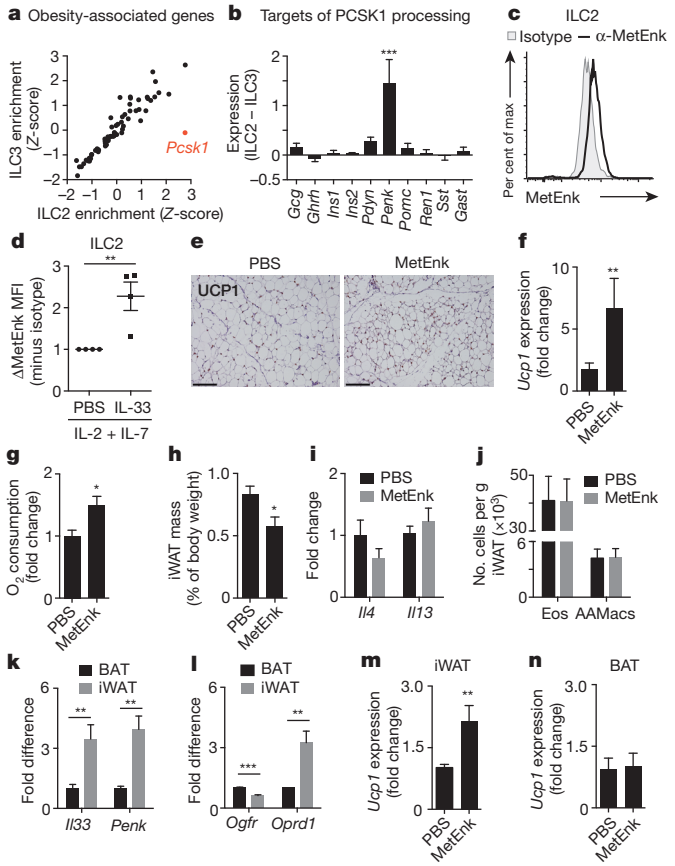


Figure 4 | ILC2s produce methionine-enkephalin, a peptide that promotes beige fat formation. **a**, Gene expression enrichment analyses of 69 obesity-associated genes in ILC2s (*x* axis, $n = 4$) versus ILC3s (*y* axis, $n = 4$). Genes significantly enriched in one cell type but not the other are red. **b**, Differential expression of PCSK1 target genes in ILC2s versus ILC3s. **c**, Intracellular staining of MetEnk (black line) or rabbit IgG isotype control (shaded histogram) in ILC2s sort-purified from E-WAT and re-stimulated *in vitro* with IL-2 and IL-7 (10 ng ml^{-1}) for 4 days. **d**, MetEnk mean fluorescence intensity (MFI) in sort-purified E-WAT ILC2s re-stimulated *in vitro* with IL-2 and IL-7 (10 ng ml^{-1}) with or without IL-33 (30 ng ml^{-1}) for 4 days. Isotype control MFI for each group was subtracted before calculating relative expression. Shown are averages from 4 independent experiments, each representing pooled cells from $n = 3$-5 mice and measured in duplicate or triplicate. **e**–**j**, Wild-type mice were treated with PBS ($n = 7$) or MetEnk ($n = 9$) by subcutaneous injection (10 mg per kg body weight per day) for 5 days. **e**, Uncoupling protein 1 (UCP1) immunohistochemistry (IHC) in inguinal white adipose tissue (iWAT). Scale bars, 100 μm. **f**, iWAT *Ucp1* expression; **g**, iWAT oxygen consumption; **h**, iWAT relative mass; **i**, iWAT *Il4* and *Il13* expression and **j**, numbers of eosinophils (Eos, live CD45$^+$ SiglecF$^+$ SSC^{hi}) and alternatively activated macrophages (AAMacs, live CD45$^+$ SiglecF$^-$ F4/80$^+$ CD206$^+$) per gram of iWAT. **k**, *Il33* and *Penk* mRNA and **l**, *Ogfr* and *Oprd1* mRNA in iWAT versus brown adipose tissue (BAT), $n = 8$. **m, n**, Stromal vascular fraction (SVF) cells from **m**, iWAT or **n**, BAT of 4-week-old C57BL/6 mice were differentiated into primary adipocytes for 2 days, treated with PBS or 50 μM MetEnk from days 2–8 and harvested on day 8 (iWAT: $n = 7$ PBS, $n = 8$ MetEnk; BAT: $n = 6$ PBS, $n = 6$ MetEnk). Student's *t*-test or ANOVA, $^*P < 0.05$, $^{***}P < 0.001$. Data are shown as mean ± standard error and are representative of 2–3 independent experiments. Sample sizes are biological replicates.

Ucp1 expression in cultured primary adipocytes from iWAT (Fig. 4m) but not BAT (Fig. 4n). Taken together, these results identify that ILC2s express MetEnk that can directly promote beiging of WAT (Extended Data Fig. 8).

To our knowledge, these data collectively provide the first demonstration that dysregulated ILC2 responses in WAT are a conserved feature of obesity in humans and mice and that the IL-33/ILC2 axis regulates metabolic homeostasis by eliciting beiging of white adipose tissue. Production of enkephalin peptides is a previously unrecognized effector mechanism employed by ILC2s to regulate metabolic homeostasis. From an evolutionary perspective, coupling ILC2-dependent innate immune effector functions with the maintenance of systemic metabolic homeostasis could provide a rapid, integrated multi-organ response that allows mammals to surmount multiple environmental challenges including infection, nutrient stress or changes in temperature. Given that impaired beige adipocyte function is associated with increased weight gain and obesity in mice[9,19] and that activity of brown/beige[17,30] adipose tissue is dysregulated in obese patients[20,21], targeting the IL-33/ILC2/beiging pathway could represent a new approach for treating obesity and obesity-associated diseases.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 31 March; accepted 27 November 2014.
Published online 22 December 2014.

1. Lumeng, C. N. & Saltiel, A. R. Inflammatory links between obesity and metabolic disease. *J. Clin. Invest.* **121,** 2111–2117 (2011).
2. Osborn, O. & Olefsky, J. M. The cellular and signaling networks linking the immune system and metabolism in disease. *Nature Med.* **18,** 363–374 (2012).
3. Moro, K. *et al.* Innate production of T_H2 cytokines by adipose tissue-associated c-Kit$^+$Sca-1$^+$ lymphoid cells. *Nature* **463,** 540–544 (2010).
4. Halim, T. Y. *et al.* Group 2 innate lymphoid cells are critical for the initiation of adaptive T helper 2 cell-mediated allergic lung inflammation. *Immunity* **40,** 425–435 (2014).
5. Molofsky, A. B. *et al.* Innate lymphoid type 2 cells sustain visceral adipose tissue eosinophils and alternatively activated macrophages. *J. Exp. Med.* **210,** 535–549 (2013).
6. Hams, E., Locksley, R. M., McKenzie, A. N. & Fallon, P. G. Cutting edge: IL-25 elicits innate lymphoid type 2 and type II NKT cells that regulate obesity in mice. *J. Immunol.* **191,** 5349–5353 (2013).
7. Harms, M. & Seale, P. Brown and beige fat: development, function and therapeutic potential. *Nature Med.* **19,** 1252–1263 (2013).
8. Shabalina, I. G., *et al.* UCP1 in brite/beige adipose tissue mitochondria is functionally thermogenic. *Cell Rep.* **5,** 1196–1203 (2013).
9. Cohen, P. *et al.* Ablation of PRDM16 and beige adipose causes metabolic dysfunction and a subcutaneous to visceral fat switch. *Cell* **156,** 304–316 (2014).
10. Price, A. E. *et al.* Systemically dispersed innate IL-13-expressing cells in type 2 immunity. *Proc. Natl Acad. Sci. USA* **107,** 11489–11494 (2010).
11. Neill, D. R. *et al.* Nuocytes represent a new innate effector leukocyte that mediates type-2 immunity. *Nature* **464,** 1367–1370 (2010).
12. Miller, A. M. *et al.* Interleukin-33 induces protective effects in adipose tissue inflammation during obesity in mice. *Circ. Res.* **107,** 650–658 (2010).
13. Mjösberg, J. M. *et al.* Human IL-25- and IL-33-responsive type 2 innate lymphoid cells are defined by expression of CRTH2 and CD161. *Nature Immunol.* **12,** 1055–1062 (2011).
14. Monticelli, L. A. *et al.* Innate lymphoid cells promote lung-tissue homeostasis after infection with influenza virus. *Nature Immunol.* **12,** 1045–1054 (2011).
15. Yang, Q. *et al.* T cell factor 1 is required for group 2 innate lymphoid cell generation. *Immunity* **38,** 694–704 (2013).
16. Miller, A. M. *et al.* IL-33 reduces the development of atherosclerosis. *J. Exp. Med.* **205,** 339–346 (2008).
17. Wu, J. *et al.* Beige adipocytes are a distinct type of thermogenic fat cell in mouse and human. *Cell* **150,** 366–376 (2012).
18. Rosen, E. D. & Spiegelman, B. M. What we talk about when we talk about fat. *Cell* **156,** 20–44 (2014).
19. Feldmann, H. M., Golozoubova, V., Cannon, B. & Nedergaard, J. UCP1 ablation induces obesity and abolishes diet-induced thermogenesis in mice exempt from thermal stress by living at thermoneutrality. *Cell Metab.* **9,** 203–209 (2009).
20. Carey, A. L. *et al.* Ephedrine activates brown adipose tissue in lean but not obese humans. *Diabetologia* **56,** 147–155 (2013).
21. Saito, M. *et al.* High incidence of metabolically active brown adipose tissue in healthy adult humans: effects of cold exposure and adiposity. *Diabetes* **58,** 1526–1531 (2009).
22. Qiu, Y. *et al.* Eosinophils and type 2 cytokine signaling in macrophages orchestrate development of functional beige fat. *Cell* **157,** 1292–1308 (2014).
23. Wu, D. *et al.* Eosinophils sustain adipose alternatively activated macrophages associated with glucose homeostasis. *Science* **332,** 243–247 (2011).
24. Liu, P.-S. *et al.* Reducing RIP140 expression in macrophage alters ATM infiltration, facilitates white adipose tissue browning, and prevents high-fat diet-induced insulin resistance. *Diabetes* **63,** 4021–4031 (2014).
25. Feuerer, M. *et al.* Lean, but not obese, fat is enriched for a unique population of regulatory T cells that affect metabolic parameters. *Nature Med.* **15,** 930–939 (2009).
26. McCarthy, M. I. Genomics, type 2 diabetes, and obesity. *N. Engl. J. Med.* **363,** 2339–2350 (2010).

27. Walley, A. J., Asher, J. E. & Froguel, P. The genetic contribution to non-syndromic human obesity. *Nature Rev. Genet.* **10,** 431–442 (2009).
28. Seidah, N. G., Sadr, M. S., Chretien, M. & Mbikay, M. The multifaceted proprotein convertases: their unique, redundant, complementary, and opposite functions. *J. Biol. Chem.* **288,** 21473–21481 (2013).
29. Lloyd, D. J., Bohan, S. & Gekakis, N. Obesity, hyperphagia and increased metabolic efficiency in Pc1 mutant mice. *Hum. Mol. Genet.* **15,** 1884–1893 (2006).
30. Sharp, L. Z. *et al.* Human BAT possesses molecular signatures that resemble beige/brite cells. *PLoS ONE* **7,** e49452 (2012).

Acknowledgements The authors wish to thank members of the Artis laboratory for the critical reading of this manuscript. Research in the Artis laboratory is supported by the National Institutes of Health (AI061570, AI074878, AI095466, AI095608, AI102942, and AI097333 to D.A.), the Burroughs Wellcome Fund Investigator in Pathogenesis of Infectious Disease Award (D.A.) and Crohn's & Colitis Foundation of America (D.A.). Additional funding was provided by NIH F30-AI112023 (J.R.B.), T32-AI060516 (J.R.B.), T32-AI007532 (L.A.M.), KL2-RR024132 (B.S.K.), DP5OD012116 (G.F.S.), P01AI06697 (D.L.F.), F31AG047003 (J.J.T.) and DP2OD007288 (P.S.) and by the Searle Scholars Award (P.S.). We thank M. A. Lazar for scientific and technical advice, D. E. Smith for providing *Il33*$^{-/-}$ mice, A. Goldrath for providing *Id2*$^{-/-}$ chimaeras, and A. Bhandoola for providing *Tcf7*$^{-/-}$ mice. We also thank the Mouse Phenotyping, Physiology & Metabolism Core at the Diabetes Research Center (DRC) of the Institute for Diabetes, Obesity & Metabolism (IDOM) as well as the Penn Diabetes Endocrine Research Center Grant (P30DK19525). In addition, we thank the Matthew J. Ryan Veterinary Hospital Pathology Laboratory, the Penn Microarray Facility, and the Mucosal Immunology Studies Team (MIST) of the NIH NIAID for shared expertise and resources. The authors would also like to thank the Abramson Cancer Center Flow Cytometry and Cell Sorting Resource Laboratory for technical advice and support. The ACC Flow Cytometry and Cell Sorting Shared Resource is partially supported by NCI Comprehensive Cancer Center Support Grant (no. 2-P30 CA016520). This work was supported by the NIH/NIDDK P30 Center for Molecular Studies in Digestive and Liver Diseases (P30-DK050306), its pilot grant program and scientific core facilities (Molecular Pathology and Imaging, Molecular Biology, Cell Culture and Mouse), as well as the Joint CHOP-Penn Center in Digestive, Liver and Pancreatic Medicine and its pilot grant program. In addition, we would like to acknowledge and thank the New York Organ Donor Network, the Cooperative Human Tissue Network-Eastern Division and especially the donors and their families. We apologize to colleagues whose work we were unable to quote owing to space constraints.

Author Contributions J.R.B., B.S.K., S.A.S., R.R.S., L.A.M., G.F.S., K.L., P.S. and D.A. designed and performed the research and/or provided advice and technical expertise. J.J.T. and D.L.F. provided human tissues. J.R.B. and D.A. analysed the data and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.A. (dartis@med.cornell.edu).

METHODS

Mice. C57BL/6, CD45.1$^+$ C57BL/6, *Rag1*$^{-/-}$ and *DblGata1* (Balb/c background) mice were obtained from Jackson Labs. *Rag2*$^{-/-}$, *Rag2*$^{-/-}$ $\gamma c^{-/-}$, *Il33*$^{+/+}$, Balb/c and *Il4ra*$^{-/-}$ (Balb/c background) mice were obtained from Taconic. *Il33*$^{-/-}$ mice were provided by Amgen Inc. via Taconic. *Id2*$^{-/-}$ bone marrow chimaeras[14] and *Tcf7*$^{-/-}$ mice[15] were generated as described previously. Unless otherwise noted, all mice were on a C57BL/6 background. All mice were males and had *ad libitum* access to food and water and were maintained in a specific-pathogen free facility with a 12 h:12 h light:dark cycle. Animals were randomly assigned to groups of $n = 3$–5 mice per group per experiment, and at least two independent experiments were performed throughout. In all *in vivo* experiments, a single technical replicate per mouse was performed except in glucose homeostasis tests described below, in which 2–4 technical replicates were performed per mouse for each time point. For all mRNA analyses, biological replicates were measured in duplicate or triplicate. For all *in vitro* experiments, 2–3 technical replicates were performed in each independent experiment. Sample sizes in each independent experiment were selected to have power of at least 90% using published sample size/power formulas[31]. Studies were not blinded. All experiments were carried out under the guidelines of the Institutional Animal Care and Use Committee at the University of Pennsylvania.

Human samples. Subcutaneous white adipose tissue (S-WAT) from the abdominal region was obtained from human donors via the New York Human Organ Donor Network (NYODN) and via the Cooperative Human Tissue Network (CHTN) Eastern Division, University of Pennsylvania. Donor characteristics are summarized in Extended Data Table 1. NYODN samples were from recently deceased organ donors at the time of organ acquisition for clinical transplantation through an approved research protocol and MTA with the NYODN. All NYODN donors were free of cancer and were hepatitis B, hepatitis C and human immunodeficiency virus-negative. Tissues were collected after the donor organs were flushed with cold preservation solution and clinical procurement process was completed. Samples from CHTN were collected from non-deceased adults undergoing panniculectomies, and were harvested from discarded connective tissue by CHTN staff. All human samples from NYODN and CHTN were stored in DMEM on ice or at 4 °C for 24–48 h before processing. Donors were defined as non-obese if their body mass index (BMI) was $<30.0\ kg\ m^{-2}$ ($n = 7$) or obese if their BMI was $\geq 30.0\ kg\ m^{-2}$ ($n = 7$). Sample sizes per group were selected to have power >95% using published sample size/power formulas[31]. There were no differences in the proportion of donors from NYODN or CHTN between non-obese and obese groups (Extended Data Table 1). ILC2 frequencies were also compared for all characteristics shown in Extended Data Table 1, and those characteristics that had a P value < 0.10 were interrogated to test whether they could explain the differences in ILC2 frequencies observed between non-obese versus obese donors. The human samples from NYODN do not qualify as 'human subjects' research, as confirmed by the Columbia University IRB, and the human samples from CHTN were de-identified and were not obtained for the specific purpose of these studies and therefore are not considered 'human subjects' research.

Diet-induced obesity. Where indicated, mice were fed a control diet (CD, 10% kcal fat, Research Diets, New Brunswick, New Jersey) or high fat diet (HFD, 45% or 60% kcal fat as indicated, Research Diets) for the indicated period of time starting at 6–8 weeks of age. CD and HFD were gamma-irradiated (10–20 kGy). In all experiments that did not employ HFD or CD, mice were fed a standard autoclavable rodent chow (5% kcal fat, 5010, Lab Diets, St. Louis, Missouri).

***In vivo* cytokine and enkephalin peptide treatments.** Mice were administered 12.5 μg per kg body weight carrier-free recombinant murine IL-33 (rmIL-33, R&D Systems, Minneapolis, Minnesota) in sterile phosphate buffered saline (PBS) by intraperitoneal (i.p.) injection for 7 days at the indicated dose. In HFD studies, mice were treated with 12.5 μg per kg body weight recombinant murine IL-33 or PBS once every 4 days by i.p. injection. In some studies, mice were treated with a previously reported[32] dose of 10 mg per kg body weight [Met^5]-enkephalin acetate salt hydrate (MetEnk, amino acid sequence YGGFM, ≥ 95.0% purity by HPLC, Sigma Aldrich, St. Louis, MO) in PBS or with PBS alone by bilateral subcutaneous injection near the iWAT daily for 5 days (approximately 200 μl per side). MetEnk or vehicle injections were performed under isoflurane anaesthesia.

Sort-purification and transfer of ILC2s. E-WAT was harvested from male CD45.1$^+$ C57BL/6 mice that received daily injections of rmIL-33 (12.5 μg per kg body weight) for 7 days by intraperitoneal injection. Live CD45$^+$ Lin$^-$ CD25$^+$ IL-33R$^+$ ILC2s were sort-purified using an Aria Cell Sorter (BD) to ≥ 98% purity, and 10^5 ILC2s were immediately transferred to the indicated recipient mice by intraperitoneal injection (5×10^4 cells) and by subcutaneous injection near iWAT (5×10^4 cells split evenly for bilateral injections). Daily transfers were performed for 4 consecutive days, and tissues were harvested on day 5. In ILC2 reconstitution experiments involving *Rag2*$^{-/-}$ $\gamma c^{-/-}$ recipient mice, 10^5 ILC2s were transferred by a single intraperitoneal injection, and the next day mice were treated with PBS or rmIL-33 (12.5 μg per kg body weight) by daily intraperitoneal injection for 7 days.

***In vivo* metabolic phenotyping.** Mice were single-housed in an OxyMax Comprehensive Laboratory Animal Monitoring System (CLAMS, Columbus Instruments, Columbus, Ohio) for 24 h. Mice were acclimated to the CLAMS cages for 24 h before measurements commenced. Fat mass and adiposity were measured by ^{1}H-nuclear magnetic resonance (NMR) spectroscopy. For glucose tolerance tests, mice were fasted overnight for 14–16 h and injected with 2 g per kg body weight D-glucose by i.p. injection. Blood glucose values were measured just before injection (time 0) and at 20, 40, 60, 90 and 120 min post-injection. For insulin tolerance tests, mice were fasted for 4–6 h and then injected with bovine insulin (0.5 U per kg body weight). Blood glucose values were measured just before injection (time 0) and at 20, 40 and 60 min post-injection. To measure fasting blood glucose and insulin concentrations, mice were fasted overnight for 14–16 h, and blood glucose values were measured followed by collection of approximately 20–30 μl blood for serum insulin concentration determination using the Ultra Sensitive Mouse Insulin ELISA Kit (Crystal Chem). Homeostatic model assessment of insulin resistance (HOMA-IR) index values were calculated as described previously[33]. All blood glucose measurements were performed using FreeStyle Lite handheld glucometer (Abbot) in duplicate or triplicate.

Histologic analysis. Tissues were fixed in 4% paraformaldehyde in PBS for at least 48 h at 4 °C and embedded in paraffin before cutting 5-μm sections and staining with haematoxylin and eosin (H&E) or performing immunohistochemistry (IHC) with rabbit anti-UCP1 antibody (Abcam, ab10983). For IHC, rehydrated sections were microwaved in 10 mM citric acid buffer (pH 6.0) for antigen retrieval, and endogenous peroxidases were quenched with 3% hydrogen peroxide. Sections were blocked with Avidin D, biotin and protein blocking agent in sequential order followed by application of the anti-UCP1 antibody (1:500). A biotinylated anti-rabbit antibody was used as a secondary antibody. Horseradish peroxidase-conjugated ABC reagent was applied, and then DAB reagent was used to develop the signal before counterstaining in haematoxylin and dehydrating the sections in preparation for mounting. Stained sections were visualized and photographed using a Nikon E600 bright field microscope.

Adipocyte area quantification. Inguinal white adipose tissue (iWAT) sections were H&E stained and imaged at ×40 magnification. White adipocyte area was calculated using ImageJ software by drawing ellipses circumscribing white adipocytes. The scale was set to 8 pixels per μm based on the pixel length of a 100-μm scale bar at ×40 magnification. Two to three images, each from a different area of a given sample, were captured per animal. Adipocyte area was measured in 10–20 adipocytes per image (25–40 adipocytes per mouse) and averaged on a per-mouse basis.

Isolation of immune cells and flow cytometry. Murine epididymal white adipose tissue (E-WAT), inguinal WAT (iWAT) or brown adipose tissue (BAT) or human subcutaneous abdominal WAT were harvested and digested with 0.1% collagenase type II (Sigma-Aldrich, USA) at 37 °C with shaking at 200 r.p.m. for 60–90 min. Digested tissues were filtered through a 70-μm nylon mesh and centrifuged at 500*g* for 5 min. Floating adipocytes were removed, and the stromal vascular fraction (SVF) pellet was resuspended in red blood cell lysis buffer (ACK RBC Lysis Buffer). Recovered cells were washed and stained with live/dead stain (Molecular Probes) followed by standard surface staining for flow cytometric analysis with fluorochrome-conjugated antibodies. Murine cells were stained with combinations of the following antibodies: anti-mouse CD45-eFluor 605NC (clone 30-F11), CD45.1-eFluor 450 (A20), CD45.2-AlexaFluor 700 (104), F4/80-eFluor 450 (BM8), CD3e-PerCP-Cy5.5 (145-2C11), CD5-PerCP-Cy5.5 (53-7.3), CD19-PerCP-Cy5.5 (1D3), NK1.1-PerCP-Cy5.5 (PK136), CD11c-PerCP-Cy5.5 (N418), FcεRIα-FITC (MAR-1), Foxp3-FITC (FJK-16 s), GATA-3-PE (TWAJ) and CD25-PE-Cy7 (clone PC61.5) from eBioscience (San Diego, CA); CD11b-PE-Texas Red (M1/70.15) from Life Technologies (Grand Island, NY); CD90.2-Alexa Fluor 700 (30-H12) and CD4-Brilliant Violet-650 (RM4-5) from BioLegend (San Diego, CA); SiglecF-PE (E50-2440) and CD3e-PE-CF594 (145-2C11) from BD Biosciences (San Jose, CA); IL-33R-biotin (T1/S2, clone DJ8) from MD Bioproducts (St. Paul, MN); CD206-Alexa Fluor 647 (MR5D3) from AbD Serotec (Raleigh, NC); and Streptavidin-APC from eBioscience. Foxp3, GATA-3 and CD206 staining was performed following fixation and permeabilization with the Foxp3 Staining Buffer Set (eBioscience). Human cells were stained with anti-human GATA-3-PE (TWAJ), TCRαβ-PerCP-Cy5.5 (IP26), CD5-PerCP-Cyanin5.5 (L17F12), CD19-Alexa Fluor 700 (HIB19), CD11c-Alexa Fluor 700 (3.9), CD127-eFluor 780 (eBioRDR5), CD45-eFluor 605NC (HI30), FcεRIα-biotin (AER-37) and Streptavidin-eFluor 650NC from eBioscience; CD56-Alexa Fluor 700 (B159), CD16 (3G8), CD3 (SP34-2) and CD25-PE-Cy7 (M-A251) from BD Pharmingen; CD11b-PE-Texas Red (M1/70.15) from Life Technologies; and ST2L-FITC from MD Bioproducts. Stained cells were acquired on a BD LSRII flow cytometer (BD Biosciences), and data were analysed using FlowJo software version 9.6.4 (Tree Star, Inc.).

Intracellular cytokine analysis. To examine ILC2 effector cytokine production, single-cell suspensions of E-WAT or iWAT SVF were stimulated for 4 h *ex vivo* with phorbol 12-myristate 13-acetate (PMA) ($100\ ng\ ml^{-1}$) and ionomycin ($1\ ng\ ml^{-1}$) in the presence of brefeldin A ($10\ \mu g\ ml^{-1}$) (all from Sigma-Aldrich) in a 37 °C incubator

(5% CO_2). Cells were then surface-stained, fixed and permeabilized using Cyto Fix/Perm (BD Pharmingen) according to manufacturer's instructions before intracellular staining for IL-5 (APC-IL-5, clone TRFK5, eBioscience) and IL-13 (PE-IL-13, eBio13A, eBioscience). Monensin (1:1500) was also used for intracellular staining with rabbit anti-mouse MetEnk (bs-1759R, Bioss USA, Woburn, MA) or rabbit anti-mouse IgG (Isotype control, Bioss USA) followed by staining with goat anti-rabbit PE (sc-3739, Santa Cruz Biotechnology, Dallas, TX).

Real-time PCR. Adipose tissues were snap-frozen in TRIzol (Invitrogen) and homogenized using a Tissue Lyser (Qiagen). RNA was isolated from the aqueous phase using the RNeasy kit (Qiagen) in accordance with the manufacturer's instructions. cDNA was synthesized from 1.0 μg RNA using Superscript II Reverse Transcriptase (Invitrogen) and oligo(dT) (Invitrogen). Real-time PCR was performed using SYBR Green technology (Applied Biosystems) with previously published primer sequences for murine *Ucp1*[17] and Qiagen QuantiTect real-time PCR primers for β-actin, *Il33*, *Penk*, *Oprd1 and Ogfr*. Reactions were run on the 7500 Fast Real-Time PCR System (Applied Biosystems) or the QuantStudio 6 Flex Real-Time PCR System (Applied Biosystems). Results were normalized to the housekeeping gene β-actin, and the $\Delta\Delta C_t$ method was employed for all real-time PCR analyses.

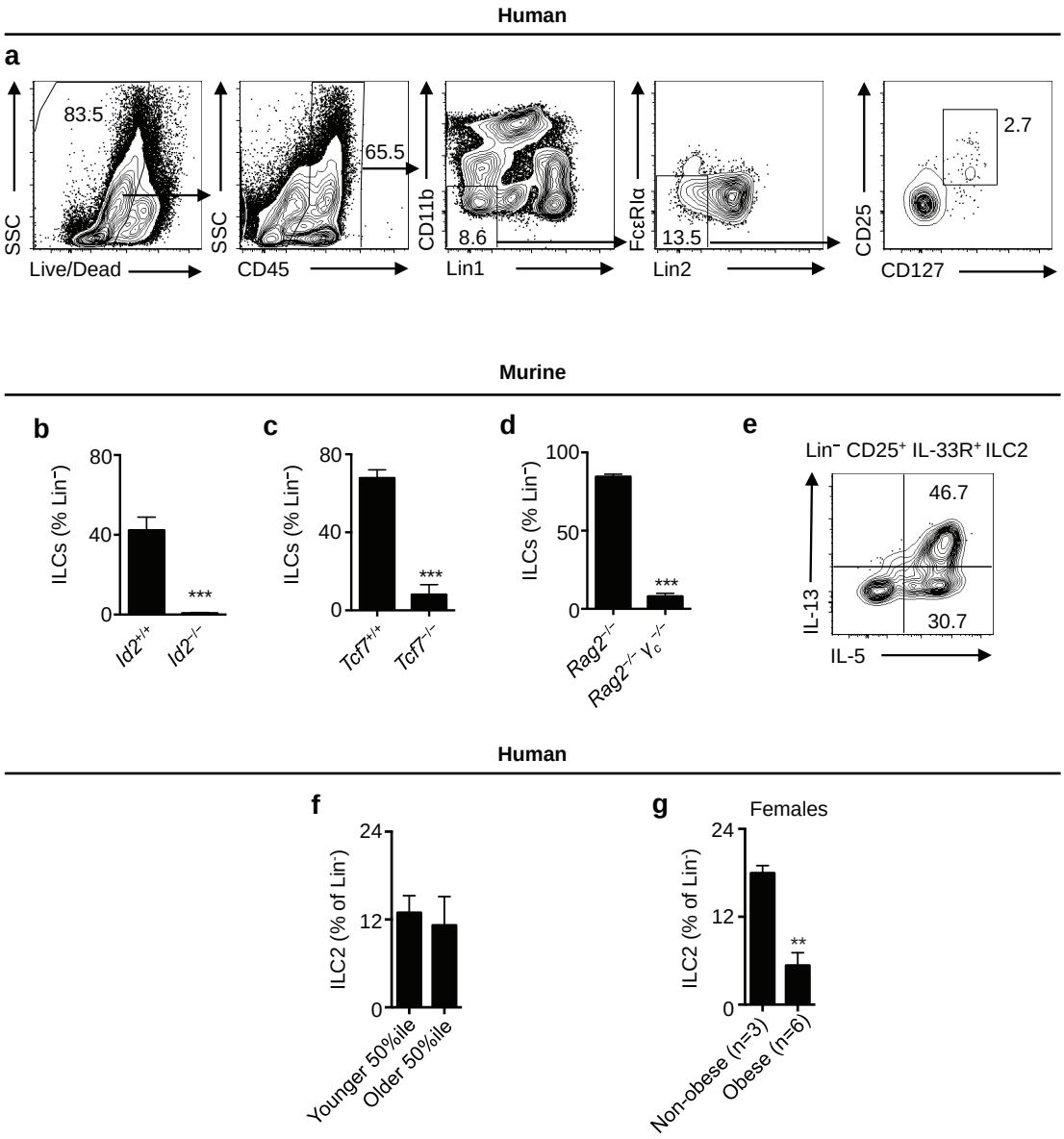
Microarrays and ILC2 versus ILC3 gene enrichment analyses. Microarray analyses (~25,000 genes) were performed using previously published microarray data sets (GEO GSE46468)[14]. In brief, Lin^- $CD90^+$ $CD25^+$ IL-$33R^+$ ILC2s from the lung (4 biological replicates each comprising 6 pooled lungs) and Lin^- $CD90^+$ $CD25^+$ $CD4^+$ ILC3s from the spleen (4 biological replicates each comprising 10 pooled spleens) were sorted using a FACS Aria (BD) directly into TRIzol LS (Invitrogen) at a purity of >97% ($1.5–2.0 \times 10^4$ cells per replicate). mRNA was isolated, amplified, labelled and hybridized to Affymetrix GeneChip (Mouse Gene 1.0 ST) as described previously[14]. Gene expression *Z*-scores were calculated for each of 69 obesity-associated genes in ILC2s or ILC3s (see Extended Data Table 2 for a complete list of genes). Genes that were significantly enriched compared to the average gene expression level of the entire microarray data set) in one cell population ($Z > 2.20$) but not the other were considered to be differentially enriched in that cell population. Bonferroni correction (alpha = 0.05, $k = 69$) was applied for microarray analyses to account for multiple testing.

Tissue oxygen consumption. A ~20 mg biopsy of iWAT was isolated from directly below the lymph node and minced in PBS containing 2% BSA, 1.1 mM sodium pyruvate and 25 mM glucose. Samples were placed in an MT200A Respirometer Cell (Strathkelvin), and oxygen consumption was measured for approximately 5 min. Oxygen consumption rates were normalized to minced tissue weight.

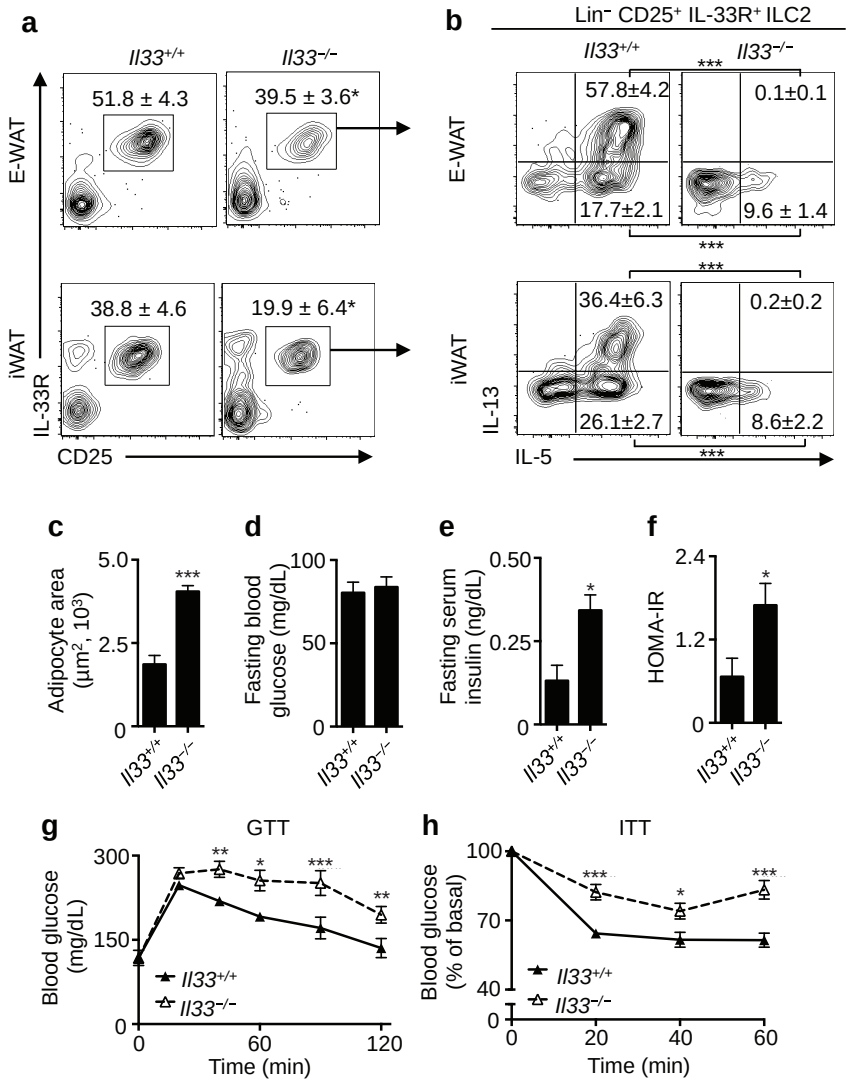
Primary adipocyte culture. iWAT or BAT was dissected from 4 week-old C57BL/6 mice ($n = 5$ per experiment, pooled) and digested as described above. Stromal vascular fraction (SVF) cells were plated in 12-well CellBind plates, and adherent cells were grown to confluence. Cells were differentiated into adipocytes as previously described[34]. Briefly, cells were cultured for 2 days with 850 nM insulin, 1 nM 3,3′,5-triiodo-L-thyronine (T_3), 1 μM rosiglitazone, 125 nM indomethacin (125 μM for BAT primary adipocytes), 0.5 mM isobutylmethylxanthine (IBMX) and 1 μM dexamethasone in adipocyte culture media (DMEM:F12 [50:50] supplemented with 10% heat-inactivated FBS, penicillin, streptomycin and L-glutamine). Cells were then maintained in adipocyte culture media supplemented with 850 nM insulin and 1 nM T_3 with either PBS or 50 μM MetEnk for 6 days, with fresh media replacement every 2 days. Cells were harvested on day 8 in TRIzol.

Statistical analyses. Data are expressed as mean ± standard error of the mean (s.e.m.). Statistical significance was determined for normally-distributed data by using the two-tailed Student's *t* test or a one-way or two-way analysis of variance (ANOVA) followed by Sidak or Tukey post-hoc tests. If variance differed between groups, the appropriate statistical correction was applied (for example, Welch's correction). Correlation analyses were conducted using Pearson linear regression. Proportions among human samples were compared by Chi-squared tests. Significance was set at $P < 0.05$. Statistical analyses were performed with Prism 6 (GraphPad Software, Inc.) or SPSS Statistics version 22 (IBM).

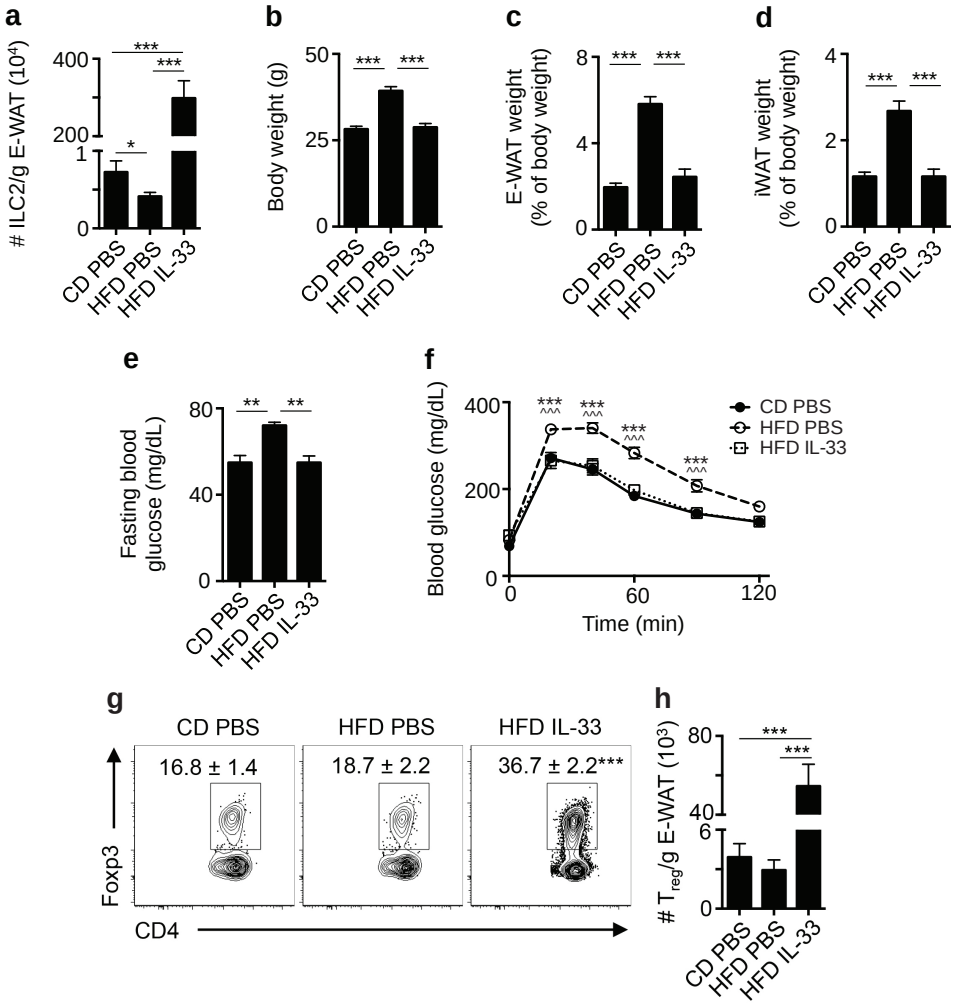
31. Brestoff, J. R. & Van den Broeck, J. in *Epidemiology: Principles and Practical Guidelines* (eds Van den Broeck, J. & Brestoff, J. R.) pp. 137–155 (Springer, 2013).
32. Zagon, I. S., Rahn, K. A., Bonneau, R. H., Turel, A. P. & McLaughlin, P. J. Opioid growth factor suppresses expression of experimental autoimmune encephalomyelitis. *Brain Res.* **1310,** 154–161 (2010).
33. Matthews, D. R. *et al.* Homeostasis model assessment: insulin resistance and beta-cell function from fasting plasma glucose and insulin concentrations in man. *Diabetologia* **28,** 412–419 (1985).
34. Seale, P. *et al.* Prdm16 determines the thermogenic program of subcutaneous white adipose tissue in mice. *J. Clin. Invest.* **121,** 96–105 (2011).



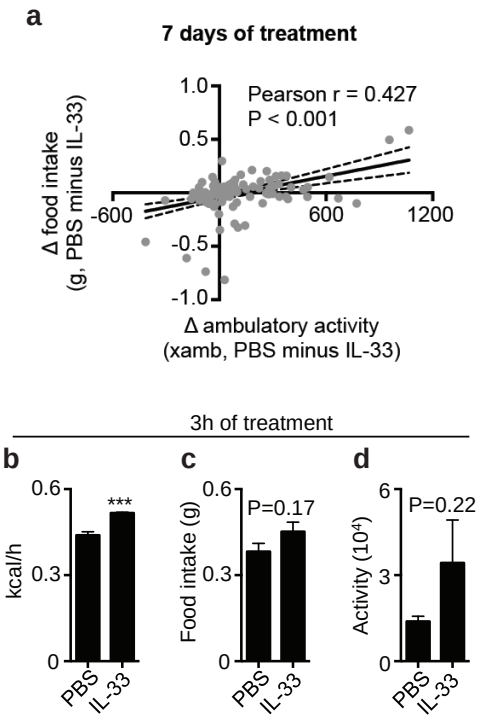
Extended Data Figure 1 | Identification of human innate lymphoid cell (ILCs) in WAT and developmental and functional characterization of murine ILCs in WAT. **a**, Gating strategy to identify human ILCs. Stromal vascular fraction (SVF) cells from human abdominal subcutaneous white adipose tissues (WAT) were isolated and subjected to flow cytometric analyses. First plot pre-gated on singlets. Lineage cocktail 1 (Lin1): CD3, CD5, TCRαβ. Lineage cocktail 2 (Lin2): CD19, CD56, CD11c, CD16. ILCs are identified as Lin-negative cells that are $CD25^+$ $CD127^+$. Plots shown are from an obese donor. **b**–**e**, SVF cells from murine epididymal (E)-WAT were isolated and subjected to flow cytometric analyses. ILCs were defined as live $CD45^+$ Lin^- $CD25^+$ $CD127^+$ cells. The lineage (Lin) cocktail included CD3, CD5, CD19, NK1.1, CD11c, CD11b and FcεRIα. Comparison of Lin^- $CD25^+$ $CD127^+$ cells in E-WAT of **b**, $Id2^{+/+}$ versus $Id2^{-/-}$ bone marrow chimaeras, **c**, $Tcf7^{+/+}$ versus $Tcf7^{-/-}$ mice and **d**, $Rag2^{-/-}$ versus $Rag2^{-/-}$ $\gamma_c{}^{-/-}$ mice. $n = 3$–8 mice per group from 2 independent experiments. **e**, E-WAT SVF cells from C57BL/6 mice were treated with PMA (100 ng ml^{-1}) and ionomycin (1 μg ml^{-1}) in the presence of Brefeldin A (10 μg ml^{-1}) for 4 h and stained for ILCs. Live $CD45^+$ Lin^- $CD25^+$ $CD127^+$ cells were pre-gated, and IL-5 and IL-13 protein levels were assessed. Plot shown is representative of $n = 12$ mice from 3 independent experiments. **f**, Human WAT ILC2 frequencies were compared in the 7 youngest donors (36.0 ± 3.5 years old) versus the 7 oldest donors (55.9 ± 1.9 years old). **g**, Human WAT ILC2 frequencies in female non-obese donors with body mass index (BMI) < 30.0 kg m^{-2} versus female obese donors with BMI ≥ 30.0 kg m^{-2}. Student's *t*-test. $^{**}P < 0.01$, $^{***}P < 0.001$. Data are shown as mean ± standard error and are representative of 2 independent experiments. Sample sizes are biological replicates.



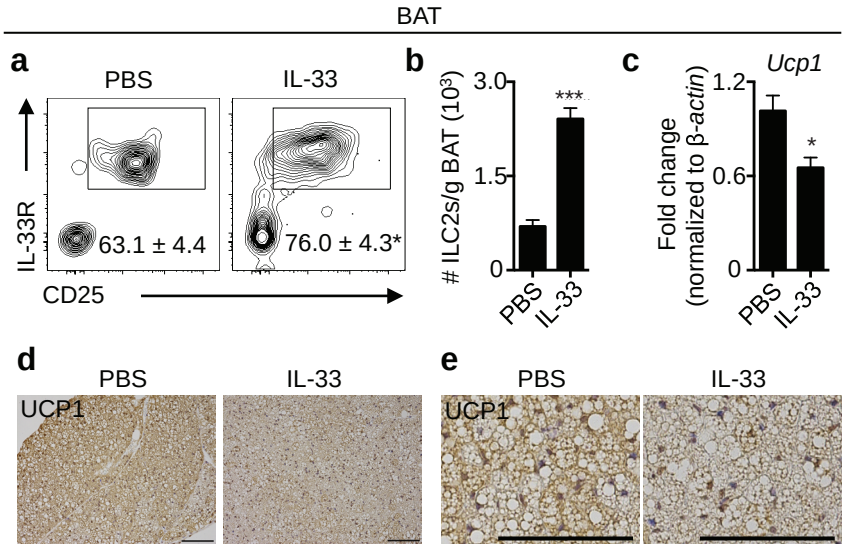
Extended Data Figure 2 | IL-33-deficient mice exhibit dysregulated group 2 innate lymphoid cells (ILC2s) in association with increased adipocyte size and impaired glucose homeostasis. *Il33*$^{+/+}$ ($n = 6$) or *Il33*$^{-/-}$ ($n = 5$) mice were fed a low-fat diet (10% kcal fat) for 12 weeks starting at 7 weeks of age. **a**, Representative plots and frequencies of live CD45^{+} Lin^{-} CD25^{+} IL-33R^{+} ILC2s in epididymal (E)-WAT (data are from Fig. 2a) and iWAT. Plots pre-gated on CD45^{+} Lin^{-} cells that lack CD3, CD5, CD19, NK1.1, CD11c, CD11b and FcεRIα. **b**, Frequencies of IL-5^{+} IL-13^{-} and IL-5^{+} IL-13^{+} ILC2s in E-WAT and iWAT of wild-type and IL-33-deficient mice. E-WAT stromal vascular fraction cells were treated with PMA (100 ng ml^{-1}) and ionomycin (1 μg ml^{-1}) in the presence of brefeldin A (10 μg ml^{-1}) for 4 h before staining for ILC2s and intracellular cytokines. Pre-gated on CD45^{+} Lin^{-} CD25^{+} IL33R^{+} ILC2s. **c**, Inguinal white adipose tissue (iWAT) sections were haematoxylin and eosin stained and imaged at ×40 magnification. Adipocyte area was calculated from 25–40 adipocytes total from 2–3 images per mouse. **d**, 16-h fasting blood glucose concentrations. **e**, 16-h fasting serum insulin concentrations. **f**, Homeostatic model assessment of insulin resistance (HOMA-IR) index values. **g**, Glucose tolerance test (GTT) with 2 g per kg body weight glucose following a 16-h fast. **h**, Insulin tolerance test (ITT) with 0.5 U per kg body weight insulin following a 5-h fast. For panels **a**–**f**, groups were compared using Student's *t*-test, $^{*}P < 0.05$, $^{***}P < 0.001$. For panels **g**–**h**, a two-way ANOVA with repeated measures was performed followed by Tukey post-hoc test. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$. Data shown are from a single cohort and are representative of 2 independent experiment. Sample sizes are biological replicates.



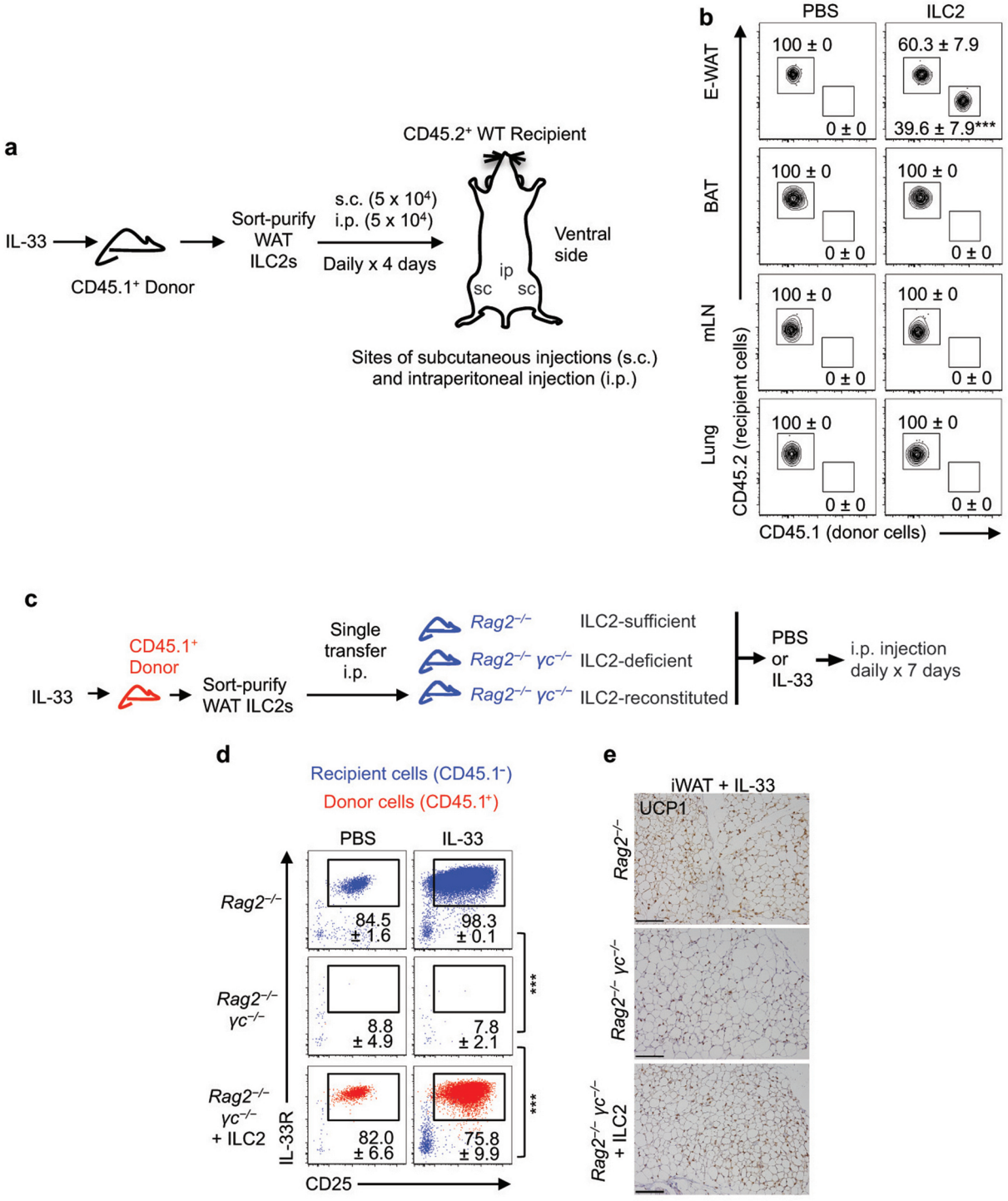
Extended Data Figure 3 | IL-33 increases E-WAT ILC2s and regulatory T cells (T_{regs}) and abrogates the development of obesity and glucose intolerance in mice fed a high-fat diet (HFD). Male C57BL/6 mice were placed on a control diet (CD) or HFD (60% kcal fat) at age 8 weeks. On the first day of feeding, CD mice were treated with PBS and HFD mice were treated with PBS or recombinant murine (rm)IL-33 (12.5 μg per kg body weight) once every 4 days by intraperitoneal injection for 4 weeks. **a**, E-WAT ILC2 numbers per gram of adipose, **b**, body weight, **c**, relative E-WAT weight and **d**, relative iWAT weight at week 4. **e**, 16-h fasting blood glucose concentrations and **f**, glucose tolerance testing during week 3. **g**, Frequencies and representative plots of E-WAT T_{regs} defined as live $CD45^+$ $CD3^+$ $CD4^+$ $Foxp3^+$ cells. Plots are gated on live $CD45^+$ $CD3^+$ $CD4^+$ cells, and numbers are the percentage of $CD4^+$ T cells that are $Foxp3^+$ T_{regs}. **h**, Numbers of T_{reg} cells per gram of adipose. All panels include $n = 10$ mice per group from 2 independent cohorts, except panel A which includes $n = 16$ CD PBS and $n = 18$ HFD PBS from 4 independent cohorts. **a**–**e**, One-way ANOVA with Tukey post-hoc test, $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$. **f**, Two-way ANOVA with repeated measures, $^{***}P < 0.001$ comparing CD PBS versus HFD PBS, $^{\wedge\wedge\wedge}P < 0.001$ comparing HFD PBS versus HFD IL-33. Data are shown as mean ± standard error. Sample sizes are biological replicates.



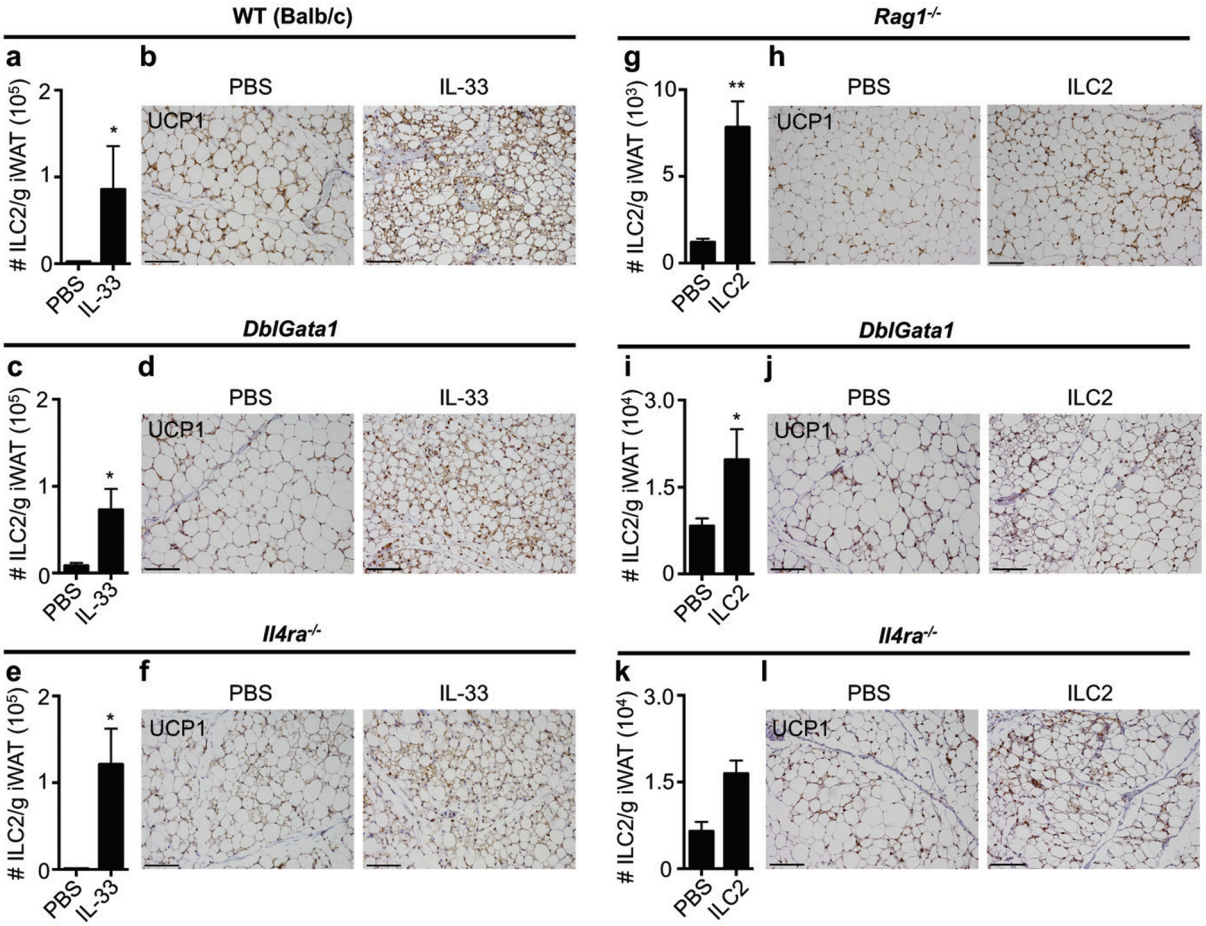
Extended Data Figure 4 | Decreased ambulatory activity may limit hyperphagia following IL-33 treatment, but IL-33 does not appear to have direct suppressive effects on food intake or ambulatory activity. **a**, Male C57BL/6 mice were treated with PBS or recombinant murine (rm)IL-33 (12.5 μg per kg body weight) daily for 7 days (PBS $n = 10$, rmIL-33, $n = 12$). Over a 24 h period between days 6 and 7, food intake and ambulatory activity were measured over 15-min intervals. The average difference in food intake or ambulatory activity between PBS- and rmIL-33-treated mice was calculated for each 15-min interval, and the differences in food intake and ambulatory activity were related by linear regression. Solid line, best-fit line. Dashed curves, upper and lower 95% confidence intervals around the best-fit line. Data are shown as mean differences for each interval and are representative of 2 independent experiments. **b**–**d**, Male C57BL/6 mice were treated with PBS or recombinant murine (rm)IL-33 (12.5 μg per kg body weight) once and monitored for the first 3 h post-treatment using CLAMS cages ($n = 4$ per group). **b**, Energy expenditure, **c**, food intake and **d**, ambulatory activity (beam breaks) were measured over of the first 3 h post-treatment. Student's *t*-test. $^{***}P < 0.001$. Data are shown as mean ± standard error and are representative of 1 independent experiment. Sample sizes are biological replicates.



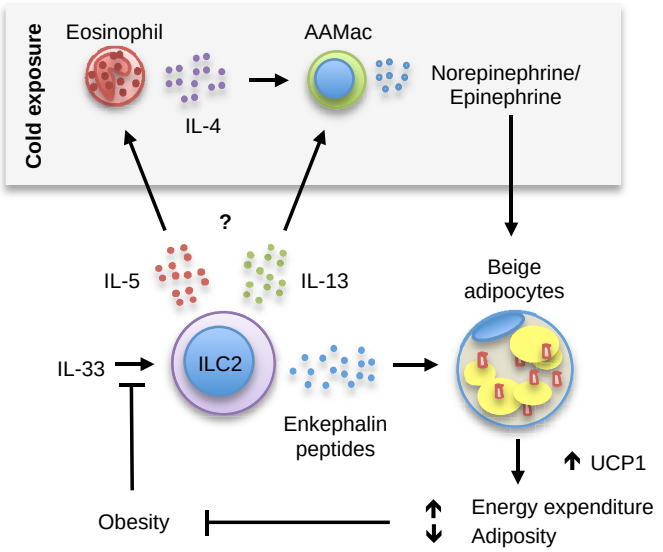
Extended Data Figure 5 | Brown adipose tissue (BAT) contains Lin$^-$ CD25$^+$ IL-33R$^+$ ILC2s that expand in response to IL-33 in association with decreased *Ucp1* expression. C57BL/6 male mice (10 weeks old) were treated with PBS ($n = 8$) or IL-33 (12.5 μg per kg body weight, $n = 8$) daily by intraperitoneal injection for 7 days. **a**, Representative plots and frequencies of Lin$^-$ CD25$^+$ IL-33R$^+$ ILC2s in interscapular BAT. Gated on live CD45$^+$ Lin$^-$ cells. **b**, Numbers of ILC2s per gram of BAT. **c**, *Ucp1* expression in BAT by real-time PCR. **d**, UCP1 immunohistochemistry of BAT at ×10 magnification. Scale bars, 100 μm. **e**, ×40 magnification of **d**. Scale bars, 100 μm. Student's *t*-test, *P < 0.05, ***$P < 0.001$. Data are shown as mean ± standard error and are representative of 2 independent experiments. Sample sizes are biological replicates.



Extended Data Figure 6 | ILC2s from E-WAT accumulate in white adipose tissue of recipient mice and expand in response to IL-33 to promote beiging. **a**, Experimental design for panels **a**, **b**. Live CD45$^+$ Lin$^-$ CD25$^+$ IL-33R$^+$ ILC2s were sort-purified from E-WAT of CD45.1$^+$ mice treated with 12.5 μg per kg body weight recombinant murine (rm)IL-33 daily for 7 days by intraperitoneal injection. PBS ($n = 8$) or ILC2s (1×10^5 total, $n = 8$) were transferred to CD45.2$^+$ recipient mice daily for 4 days by subcutaneous injection near iWAT (5×10^4 ILC2s, split evenly bilaterally) and intraperitoneal injection (5×10^4 ILC2s). Tissues were harvested on day 5 for analyses. **b**, Donor and recipient ILC2s in E-WAT, brown adipose tissue (BAT), mesenteric lymph nodes (mLN) and lung. iWAT ILC2 plots from this experiment are shown in main Fig. 3g. Pre-gated on Live CD45$^+$ Lin$^-$ CD25$^+$ IL-33R$^+$ ILC2s. Donor ILC2s are defined as CD45.1$^+$ CD45.2$^-$, whereas recipient ILC2s are defined as CD45.1$^-$ CD45.2$^+$. Representative plots shown. Frequencies represent percent of ILC2s that are recipient or donor cells. Student's *t*-test, ***$P < 0.001$. **c**, Experimental design for panels **c**–**e**. Sort-purified CD45.1$^+$ ILC2s ($\times 10^5$) from E-WAT of IL-33-treated mice (as described above) were transferred into *Rag2*$^{-/-}$ *γc*$^{-/-}$ recipients by a single intraperitoneal injection. ILC2-sufficient *Rag2*$^{-/-}$ mice, ILC2-deficient *Rag2*$^{-/-}$ *γc*$^{-/-}$ mice and ILC2-reconstituted *Rag2*$^{-/-}$ *γc*$^{-/-}$ mice were treated with PBS or rmIL-33 (12.5 μg per kg body weight) by intraperitoneal injection daily for 7 days. There were $n = 4$ mice per group. This experimental design corresponds to main Fig. 3l-m. **d**, Representative plots of live CD45.1$^+$ Lin$^-$ CD25$^+$ IL33R$^+$ ILC2s in E-WAT. Blue, recipient cells. Red, donor cells. Lineage cocktail includes CD3, CD5, CD19, NK1.1, CD11c, CD11b and FcεRIα. **e**, iWAT UCP1 IHC. Scale bars, 100 μm. ANOVA with Tukey post-hoc test, ***$P < 0.001$. Data are shown as mean ± standard error and are representative of 2 independent experiments. Sample sizes are biological replicates.



Extended Data Figure 7 | IL-33 treatment and ILC2 transfer can elicit beiging independently of eosinophils and IL-4Rα signalling. **a**–**f**, Wild-type (Balb/c), *DblGata1* mice that lack eosinophils or *Il4ra*$^{-/-}$ mice that have dysregulated alternatively activated macrophages (AAMacs) (both mutant strains on a Balb/c background) were treated with PBS or recombinant murine (rm)IL-33 (12.5 μg per kg body weight) daily by intraperitoneal injection for 7 days. **a**, iWAT ILC2 numbers per gram of adipose and **b**, iWAT UCP1 immunohistochemistry (IHC) in Balb/c mice (PBS, $n = 4$; rmIL-33, $n = 3$). **c**, iWAT ILC2 numbers per gram of adipose and **d**, iWAT UCP1 IHC in *DblGata1* mice (PBS, $n = 5$; rmIL-33, $n = 6$). **e**, iWAT ILC2 numbers per gram of adipose and **f**, iWAT UCP1 IHC in *Il4ra*$^{-/-}$ mice (PBS, $n = 4$; rmIL-33, $n = 6$). **g**, **h**, Live CD45^{+} Lin^{-} CD25^{+} IL-33R^{+} ILC2s were sort-purified from E-WAT of C57BL/6 mice treated with rmIL-33 (12.5 μg per kg body weight) daily for 5–7 days by intraperitoneal injection to *Rag1*$^{-/-}$ mice on a C57BL/6 background. ILC2s (1×10^5 total) were transferred to recipient mice daily for 4 days by subcutaneous injection (PBS, $n = 8$; ILC2, $n = 8$). **g**, iWAT ILC2 numbers per gram of adipose and **h**, iWAT UCP1 IHC. **i**–**l**, Live CD45^{+} Lin^{-} CD25^{+} IL-33R^{+} ILC2s were sort-purified from E-WAT of Balb/c mice treated with rmIL-33 (12.5 μg per kg body weight) daily for 5–7 days by intraperitoneal injection. ILC2s (1×10^5 total) were transferred to recipient mice daily for 4 days by subcutaneous injection. **i**, iWAT ILC2 numbers per gram of adipose and **j**, iWAT UCP1 IHC in *DblGata1* recipients (PBS, $n = 6$; ILC2, $n = 6$). **k**, iWAT ILC2 numbers per gram of adipose and **l**, iWAT UCP1 IHC in *Il4ra*$^{-/-}$ recipients (PBS, $n = 3$; ILC2, $n = 4$). Scale bars, 100 μm. Student's *t*-test, $^{*}P < 0.05$. Data are shown as mean ± standard error and are representative of 2 independent experiments. Sample sizes are biological replicates.



Extended Data Figure 8 | Summary model linking the IL-33/ILC2/MetEnk pathway to the regulation of beiging and obesity. Interleukin (IL)-33 acts on group 2 innate lymphoid cells (ILC2s) to upregulate production of the effector molecules IL-5, IL-13 and enkephalin peptides. ILC2-derived IL-5 promotes eosinophil homeostasis in WAT, and eosinophils in turn produce IL-4 to sustain alternatively activated macrophages (AAMacs) in WAT. ILC2-derived IL-13 can also promote AAMac responses. In the setting of chronic exposure to a cold environment, eosinophil-derived IL-4 stimulates AAMacs to produce catecholamines such as noradrenaline, which acts directly on beige adipocytes to upregulate uncoupling protein 1 (UCP1) expression and promote mitochondrial biogenesis. Although it remains unknown whether ILC2-derived IL-5 and IL-13 contribute to cold-stress-induced beiging, ILC2-derived enkephalin peptides can act directly on beige adipocytes to upregulate UCP1 and promote beiging. This results in increased energy expenditure and decreased adiposity that may counteract weight gain. In the setting of obesity, IL-33 expression in WAT is increased; however, WAT ILC2s are paradoxically decreased in both mice and humans, suggesting that the IL-33/ILC2 axis is dysregulated in obesity. This may impede the ability of ILC2s to contribute to the function of beige fat, resulting in a vicious cycle that promotes weight gain.

Extended Data Table 1 | Characteristics of non-obese and obese human donors

Characteristic	Non-obese (n=7)	Obese (n=7)	P-value*
Source of tissue (% CHTN/% NYODP)	29%/71%	43%/57%	P=0.43
Age	39.3 +/- 5.2	52.6 +/- 2.7	P=0.042
Sex, % female	43%	87%	P=0.094
BMI (kg/m^2)	23.5 +/- 1.4	42.6 +/- 3.9	P=0.0006
History of Type 2 diabetes	14%	43%	P=0.24
History of liver disease	0%	0%	n/a
History of cardiovascular disease	0%	13%	P=0.30

BMI, body mass index; CHTN, Cooperative Human Tissue Network; NYODP, New York Organ Donor Program.
* Proportions were compared by χ^2 tests. Continuous variables were compared by Student's *t*-test. Exact *P* values are shown.

Extended Data Table 2 | List of genes with single nucleotide polymorphisms associated with human obesity

Human gene symbol	Human gene name	Murine ortholog gene symbol	Inclusion in Microarray	
ADAMTS9	A disintegrin-like and metallopeptidase (reprolysin type) with thrombospondin type 1 motif, 9	*Adamts9*	Yes	
BCDIN3D	BCDIN3 domain containing	*Bcdin3d*	Yes	
BDNF	Brain-derived neurotrophic factor	*Bdnf*	Yes	
CADM2	Cell adhesion molecule 2	*Cadm2*	Yes	
CNR1	Cannabinoid type 1 receptor	*Cnr1*	Yes	
CPEB4	Cytoplasmic polyadenylation element binding protein 4	*Cpeb4*	Yes	
CTNNBL1	Catenin, beta like 1	*Ctnnbl1*	Yes	
DLK1	Delta-like homologue 1	*Dlk1*	Yes	
ENPP1	Ectonucleotide pyrophasphatase/phosphodiesterase 1	*Enpp1*	Yes	
ETV5	Ets variant 5	*Etv5*	Yes	
FAIM2	Fas apoptotic inhibitory molecule 2	*Faim2*	Yes	
FANCL	Fanconi anemia, complementation group L	*Fancl*	Yes	
FTO	fat mass and obesity associated	*Fto*	Yes	
GHSR	Growth hormone receptor secretogogue receptor	*Ghsr*	Yes	
GIPR	Gastric inhibitory polypeptide receptor	*Gipr*	Yes	
GNPDA2	Glucosamine-6-phosphate deaminase 2	*Gnpda2*	Yes	
GPRC5B	G protein-coupled receptor, family C, group 5, member B	*Gprc5b*	Yes	
GRB14	Growth factor receptor-bound protein 14	*Grb14*	Yes	
HMGA1	High mobility group AT-hook 1	*Hmga1*	Yes	
HMGCR	3-hydroxy-3-methylglutaryl-CoA reductase	*Hmgcr*	Yes	
HOXC13	Homeobox C13	*Hoxc13*	Yes	
ITPR2	Inositol 1,4,5-trisphosphate receptor, type 2	*Itpr2*	Yes	
KCTD15	Potassium channel tetramerization domain containing 15	*Kctd15*	Yes	
KLF7	Kruppel-like factor 7	*Klf7*	Yes	
LEP	Leptin	*Lep*	Yes	
LEPR	Leptin receptor	*Lepr*	Yes	
LMNA	Lamin A/C	*Lmna*	Yes	
LRP1B	Low density lipoprotein receptor-related protein 1B	*Lrp1b*	Yes	
LINGO2 (LRRN6C)	Leucine rich repeat and Ig domain containing 2	*Lingo2*	Yes	
LY86	Lymphocyte antigen 86	*Ly86*	Yes	
LYPLAL1	Lysophospholipase-like 1	*Lyplal1*	Yes	
MAF	v-maf avian musculoaponeurotic fibrosarcoma oncogene homolog	*Maf*	Yes	
MAP2K5	Mitogen-activated protein kinase kinase 5	*Map2k5*	Yes	
MC4R	Melanocortin 4 receptor	*Mc4r*	Yes	
MSRA	Methionine sulfoxide reductase A	*Msra*	Yes	
MTCH2	Mitochondrial carrier 2	*Mtch2*	Yes	
MTIF3	Mitochondrial translational initiation factor 3	*Mtif3*	Yes	
MTMR9	Myotubularin related protein 9	*Mtmr9*	Yes	
NAMPT	Nicotinamide phosphoribosyltransferase	*Nampt*	Yes	
NCR3	Natural cytotoxicity triggering receptor 3	*Ncr3*		No
NEGR1	Neuronal growth regulator 1	*Negr1*	Yes	
NFE2L3	Nuclear factor, erythroid 2-like 3	*Nfe2l3*	Yes	
NPC1	Niemann-Pick disease, type C1	*Npc1*	Yes	
NPY2R	Neuropeptide Y receptor Y2	*Npy2r*	Yes	
NRXN3	Neurexin 3	*Nrxn3*	Yes	
PCSK1	Proprotein convertase subtilisin/kexin type 1	*Pcsk1*	Yes	
PIGC	Phosphatidylinositol glycan anchor biosynthesis, class C	*Pigc*	Yes	
POMC	Proopiomelanocortin	*Pomc*	Yes	
PRKD1	Protein kinase D1	*Prkd1*	Yes	
PRL	Prolactin	*Prl*	Yes	
PTBP2	Polypyrimidine tract binding protein 2	*Ptbp2*	Yes	
PTER	Phosphotriesterase related	*Pter*	Yes	
RSPO3	R-spondin 3	*Rspo3*	Yes	
SDCCAG8	Serologically defined colon cancer antigen 8	*Sdccag8*	Yes	
SEC16B	SEC16 Homolog B	*Sec16b*	Yes	
SH2B1	SH2B adaptor protein 1	*Sh2b1*	Yes	
SLC39A8	Solute carrier family 39 (zinc transporter), member 8	*Slc39a8*	Yes	
SNRPN	Small nuclear ribonucleoprotein polypeptide N	*Snrpn*	Yes	
SOCS1	Suppressor of cytokine signaling 1	*Socs1*	Yes	
SOCS3	Suppressor of cytokine signaling 3	*Socs3*	Yes	
STAB1	Stabilin 1	*Stab1*	Yes	
TBC1D1	TBC1 (tre-2/USP6, BUB2, cdc16) domain family, member 1	*Tbc1d1*	Yes	
TBX15	T-box 15	*Tbx15*	Yes	
TFAP2B	Transcription factor AP-2 beta (activating enhancer binding protein 2 beta)	*Tfap2b*		No
TMEM160	Transmembrane protein 160	*Tmem160*	Yes	
TMEM18	Transmembrane protein 18	*Tmem18*	Yes	
TNNI3K	TNNI3 interacting kinase	*Tnni3k*	Yes	
TUB	Tubby bipartite transcription factor	*Tub*	Yes	
VEGFA	Vascular endothelial growth factor A	*Vegfa*	Yes	
ZNF608	Zinc finger protein 608	*Zfp608*	Yes	
ZNRF3	Zinc and ring finger 3	*Znrf3*	Yes	

Genes are derived from references 26 and 27.

LETTER

doi:10.1038/nature14035

Crystal structure of the human OX_2 orexin receptor bound to the insomnia drug suvorexant

Jie Yin[1], Juan Carlos Mobarec[2], Peter Kolb[2] & Daniel M. Rosenbaum[1]

The orexin (also known as hypocretin) G protein-coupled receptors (GPCRs) respond to orexin neuropeptides in the central nervous system to regulate sleep and other behavioural functions in humans[1]. Defects in orexin signalling are responsible for the human diseases of narcolepsy and cataplexy; inhibition of orexin receptors is an effective therapy for insomnia[2]. The human OX_2 receptor (OX_2R) belongs to the β branch of the rhodopsin family of GPCRs[3], and can bind to diverse compounds including the native agonist peptides orexin-A and orexin-B and the potent therapeutic inhibitor suvorexant[4]. Here, using lipid-mediated crystallization and protein engineering with a novel fusion chimaera, we solved the structure of the human OX_2R bound to suvorexant at 2.5 Å resolution. The structure reveals how suvorexant adopts a π-stacked horseshoe-like conformation and binds to the receptor deep in the orthosteric pocket, stabilizing a network of extracellular salt bridges and blocking transmembrane helix motions necessary for activation. Computational docking suggests how other classes of synthetic antagonists may interact with the receptor at a similar position in an analogous π-stacked fashion. Elucidation of the molecular architecture of the human OX_2R expands our understanding of peptidergic GPCR ligand recognition and will aid further efforts to modulate orexin signalling for therapeutic ends.

The orexin system modulates diverse behaviours in mammals, including sleep, arousal and feeding[1]. Orexin neurons in the lateral hypothalamus uniquely produce the 33-amino-acid orexin-A and 28-amino-acid orexin-B neuropeptides. Orexin receptors OX_1R and OX_2R, distributed throughout the central nervous system, respond to these peptides to control neuronal activity. Signals generated from the hypothalamus, the limbic system and the periphery converge on the orexin neurons, which act as central integrators of environmental cues and extend processes to many different brain centres. The orexin receptors belong to the rhodopsin family of GPCRs and relay neuropeptide binding at synapses into intracellular activation of heterotrimeric $G_{q/11}$ and $G_{i/o}$ (ref. 5). The importance of the orexin system in phasic control of sleep–wake cycles was highlighted by discoveries that disruption/deletion of orexin or OX_2R causes narcolepsy in dogs[6], mice[7] and humans[8]. As a result, a number of potent dual orexin receptor antagonists (DORAs) have been developed and tested over the past decade[4], culminating in US Food and Drug Administration (FDA) approval of the first-in-class drug suvorexant (Belsomra) for insomnia. Suvorexant binds to human OX_1R and OX_2R (hOX_1R and hOX_2R) with sub-nanomolar affinity, potently inhibits orexin receptor signalling in cell-based assays, and promotes the transition to rapid eye movement (REM) and slow wave sleep in animals and humans[2,4,9].

To understand better the molecular basis of orexin receptor ligand recognition and signalling, we sought to obtain a high-resolution X-ray crystal structure of hOX_2R. Protein engineering (fusion proteins[10] and thermostable mutants[11]) and lipid-mediated crystallization methods[12] have recently enabled the determination of the structures of GPCRs for diverse ligands such as biogenic amines, nucleotides, peptide hormones and lipids. OX_1R and OX_2R belong to the β branch of the rhodopsin family of GPCRs, which contains receptors for neuropeptides such as the tachykinins, oxytocin/vasopressin and neurotensin[3]. Crystal structures of thermostabilized mutants of the rat neurotensin receptor (NTSR1), in partially active[13] and inactive[14] conformations, constitute the only crystallographic data currently available for this physiologically important group of GPCRs. Our attempts to express and crystallize a hOX_2R–T4L fusion protein, an approach we originally developed for the β_2 adrenergic receptor (β_2AR)[10], were unsuccessful.

We therefore explored the use of alternative fusion protein partners that would help hOX_2R pack into a well-defined three-dimensional lattice. For candidate fusion partners, we searched for domains of fewer than 200 amino acids from extreme thermophiles, which had been previously crystallized and characterized by X-ray diffraction at high resolution and have amino and carboxy termini that are close (within 10 Å) in three-dimensional space. Using a construct in which the 196-amino-acid catalytic domain of *Pyrococcus abysii* glycogen synthase (PGS)[15] replaced 39 residues of the third intracellular loop (ICL3), we were able to grow microcyrstals of hOX_2R in a cholesterol-doped monoolein cubic phase (Extended Data Fig. 2) and solve the suvorexant-bound structure at 2.5 Å resolution (Extended Data Table 1). As expected, the PGS domain promotes tight packing of hOX_2R into a crystal lattice in which membrane layers containing the embedded GPCR alternate with aqueous layers containing the fusion partner (Fig. 1a).

The overall seven-transmembrane (TM) fold of hOX_2R resembles other GPCR structures (Fig. 1a, b). Despite a low sequence similarity (23% identity), the backbone root mean square deviation (r.m.s.d.) relative to the inactive-state β_2AR[10] is only 2.2 Å. The backbone r.m.s.d. compared with NTSR1 (22% identity) is 1.3 Å for the inactive-state[14] and 2.3 Å for the partial active-state conformation[13]. At the extracellular surface, residues 190–212 in the second extracellular loop (ECL2) form a β-hairpin (Fig. 1c, d) analogous to that seen in other peptide-binding GPCRs such as NTSR1 (ref. 13), the μ-opioid receptor[16] and CXCR4 (ref. 17)—this β-hairpin structure contains amino acids important for orexin binding and activation[18].

Superposition of suvorexant-bound hOX_2R with the antagonist-bound M3 muscarinic acetylcholine receptor[19], another G_q-coupled GPCR, shows a high degree of overlap between TM backbones at the intracellular surface (Fig. 1b). One difference is that the conserved 'DRY motif' on TM3, part of an inhibitory interaction network in the rhodopsin family of GPCRs[20], is 'DRWY' in hOX_2R. Residues $D151^{3.49}$ and $R152^{3.50}$ (superscripts are Ballesteros–Weinstein numbering throughout) make an intra-motif salt bridge, while $R152^{3.50}$ and $W153^{3.51}$ contact the cytoplasmic ends of TM5 and TM6 ($Q245^{5.60}$, $I246^{5.61}$ and $L306^{6.37}$). Overall, numerous hydrophobic and polar contacts bind TM5 and TM6 to the other TM domains, restricting the outward movement of these α-helices necessary for GPCR activation. The suvorexant-bound hOX_2R structure thus represents an inactive-state conformation, consistent with the efficacy profile of suvorexant as a DORA ligand.

The suvorexant-binding pocket is open to the extracellular space through a constricted solvent-accessible channel (Fig. 2a) rimmed by amino acids from the extracellular ends of TM2, TM5–7 and the ECL2 β-hairpin. A complex network of electrostatic interactions covers the

[1]Department of Biophysics, The University of Texas Southwestern Medical Center, Dallas, Texas 75390, USA. [2]Department of Pharmaceutical Chemistry, Philipps-University Marburg, 35032 Marburg, Germany.

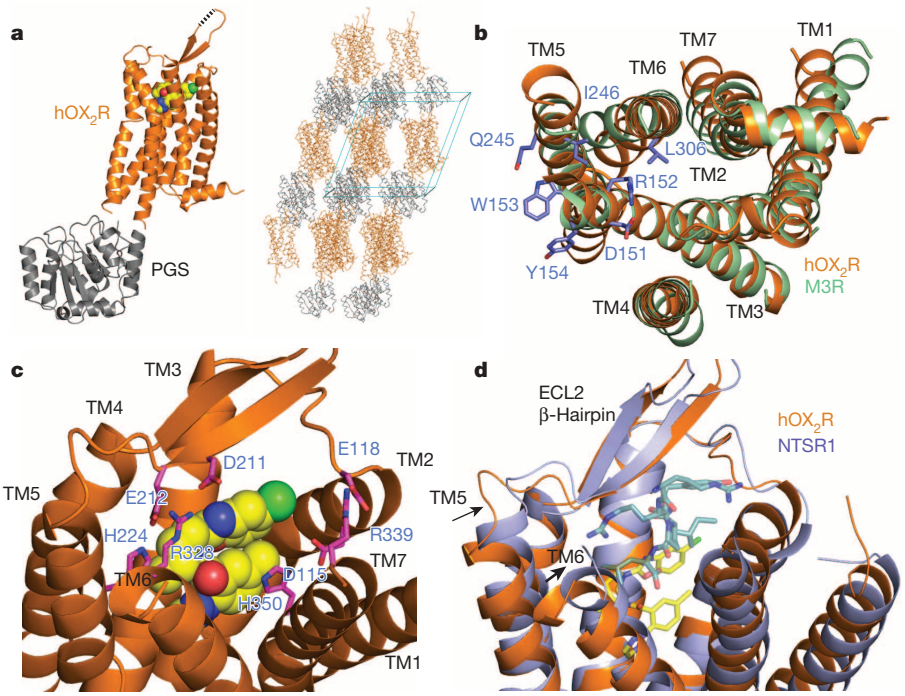


Figure 1 | Fusion protein engineering and structural features of hOX_2R. **a**, Left, global structure of the hOX_2R–PGS fusion protein. hOX_2R is represented as an orange cartoon, with the PGS domain (grey cartoon) fused at ICL3. Suvorexant is shown as spheres with yellow carbons. Dotted line represents the five amino acids that could not be modelled at the tip of the β-hairpin in ECL2. Right, packing of the hOX_2R–PGS fusion protein in the lipidic-cubic-phase-derived crystal lattice. **b**, Overlap between suvorexant-bound hOX_2R (orange cartoon) and antagonist-bound M3R[19] (green cartoon; Protein Data Bank (PDB) accession 4DAJ) at the intracellular surface. The DRWY sequence on TM3 and interacting residues on TM5 and 6 are shown as blue sticks. **c**, Salt-bridge network at the extracellular surface of hOX_2R. Residues participating in electrostatic interactions are shown as magenta sticks, with suvorexant represented as spheres with yellow carbons. ECL3 is removed for clarity. **d**, Superposition of hOX_2R (orange cartoon) and NTSR1 (blue cartoon) in a partial active-state conformation (PDB accession 4GRV)[13]. Suvorexant (yellow carbons) and the NTS_{8-13} agonist (teal carbons) are shown as transparent sticks. ECL3 is removed for clarity.

extracellular surface of the receptor, including salt bridges on both sides of the ligand entry channel ($D115^{2.65}$–$H350^{7.39}$, $E118^{2.68}$–$R339^{7.28}$, $D211^{45.51}$–$R328^{6.59}$, $E212^{45.52}$–$H224^{5.39}$) that stabilize the extracellular TM conformation (Figs 1c and 2a). A similar extracellular salt bridge in β_2AR (ECL2 to TM3) was previously shown to be a ligand-dependent switch by NMR spectroscopy[21]. Mutation of residue $D211^{45.51}$ to Ala has one of the greatest characterized deleterious effects on orexin-A potency, but has little impact on binding of some DORAs such as almorexant[18]—this amino acid is over 6 Å more extracellular than the closest suvorexant atom in the crystal structure. The difference between orexin and DORA sensitivity to $D211A^{45.51}$ suggests that modulation or competition of the extracellular salt bridges may be involved in orexin binding and activation of the receptor. In further support of this hypothesis, the neurotensin agonist peptide NTS_{8-13} present in the partially active NTSR1 structure[13] occupies a more extracellular position than suvorexant, adjacent to the β-hairpin, stabilizing a slight inward movement of TM5 and TM6 (Fig. 1d). Such inward movements of TM5 and TM6 relative to the rest of the TM bundle at the orthosteric binding pocket may be a general trigger for agonist-mediated GPCR activation, as they have also been observed for the β_2AR[22] and the M2 muscarinic acetylcholine receptor[23].

Suvorexant sterically inhibits inward motions of TM domains by lodging deep in the orthosteric site and contacting all TM α-helices except TM1 (Fig. 2b, c). The shape of suvorexant in the ligand-binding pocket resembles a horseshoe, due to a boat conformation of the diazepane ring and intramolecular π-stacking between the aromatic benzoxazole and *p*-toluamide groups (Fig. 2b, c and Extended Data Fig. 3). A similar conformation of a suvorexant analogue was previously found in small-molecule crystals and by NMR experiments in solution[24], indicating that the horseshoe probably represents a low-free-energy state of the isolated ligand. Most of the ligand contacts involve van der Waals interactions or aromatic packing, with few direct polar interactions aside from a notable hydrogen bond from $N324^{6.55}$ to suvorexant's tertiary amide carbonyl. Several water-mediated hydrogen bonds form bridges between suvorexant and polar amino acids such as $N324^{6.55}$ and $H350^{7.39}$ (Fig. 2b, c). Although the effects of mutagenesis on suvorexant affinity to hOX_2R have not been reported, certain Ala mutants appear to have a broad deleterious effect on DORA binding[18]: $W214^{5.29}$ and $Y223^{5.38}$ do not directly participate in suvorexant binding, but are critical to the structural integrity of the ECL2 β-hairpin; $F227^{5.42}$ at the base of the

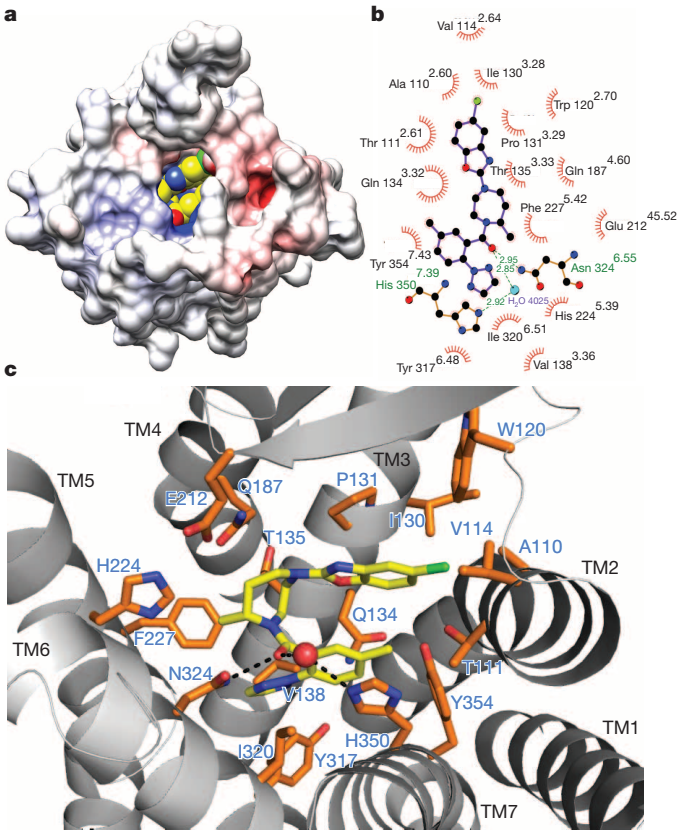


Figure 2 | Suvorexant interaction with hOX_2R. **a**, Solvent-accessible channel to the ligand-binding site. The solvent-accessible surface of the receptor is coloured according to electrostatic potential. Suvorexant is shown as spheres with yellow carbons. **b**, Two-dimensional schematic of contacts between suvorexant and the receptor. **c**, Three-dimensional interaction between suvorexant and hOX_2R, showing all residues within 4 Å of the ligand as sticks with orange carbons. Hydrogen bonds are shown as black dashes. H_2O 4025 is shown as a red sphere.

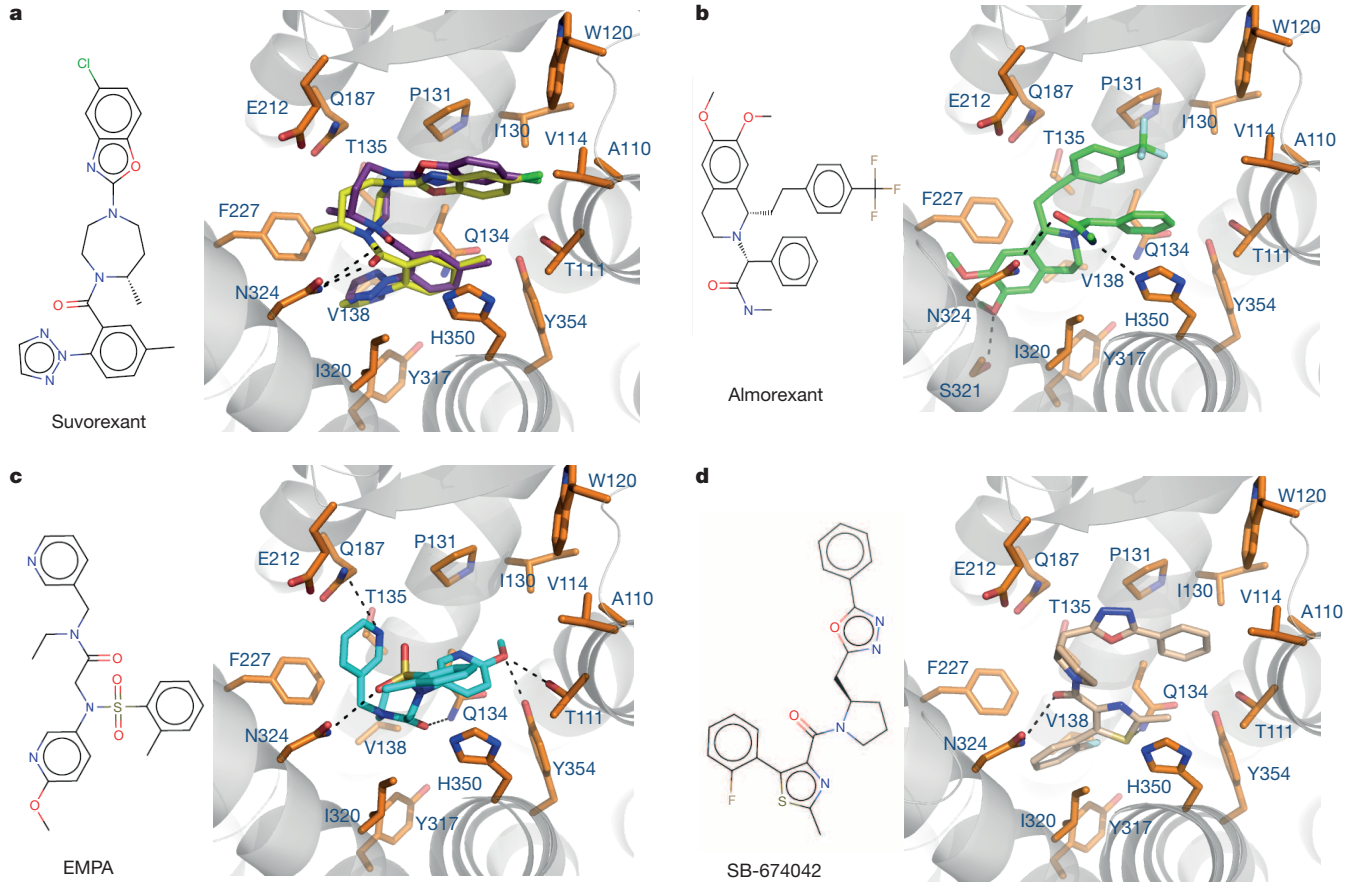


Figure 3 | Docked poses for synthetic orexin receptor antagonists. **a**, Left, chemical structure of suvorexant. Right, recapitulated binding mode of suvorexant (purple carbons) superimposed with the observed pose in the crystal structure (yellow carbons). Hydrogen bonds are shown as black dashes. **b**–**d**, Chemical structure and predicted binding mode of almorexant (green carbons) (**b**), EMPA (cyan carbons) (**c**) and SB-674042 (tan carbons) (**d**).

binding pocket packs against the methyl-diazepane ring; $Y317^{6.48}$ contacts the 1,2,3-triazole; and $H350^{7.39}$ π-stacks with the *p*-toluamide group. The residues surrounding suvorexant and the ligand entry channel are almost identical between hOX_1R and hOX_2R (Extended Data Figs 4 and 5), explaining suvorexant's ability to bind tightly and inhibit both receptors[9]. Out of 30 residues that are within 6 Å distance of suvorexant in the hOX_2R structure, only two amino acids are different compared with hOX_1R: $T111^{2.61}$ is changed to Ser and $T135^{3.33}$ is changed to Ala (overall sequence identity, 67%). This sequence conservation also implies that the 12-fold higher orexin-B affinity (and 40-fold higher potency) for hOX_2R over hOX_1R[25] is probably due to differences in interactions that are remote from the deeply membrane-embedded orthosteric binding pocket.

We have previously used computational docking methods to effectively predict interactions between a GPCR of known structure and small-molecule ligands[26]. With the newly available hOX_2R structure, we carried out molecular docking calculations (see Methods) to generate possible binding modes for three additional high-affinity orexin receptor antagonists that have chemical scaffolds distinct from suvorexant: almorexant[27], EMPA[28], and SB-674042 (ref. 29). As a control, we showed that our docking protocols were capable of accurately reproducing the interaction between suvorexant and hOX_2R in the crystal structure (Fig. 3a). Predicted poses for each of the other docked ligands establish a hydrogen bond with $N324^{6.55}$ (Fig. 3b–d), and two of the three adopt a π-stacked horseshoe-like conformation that mimics the binding of suvorexant (Fig. 3b, d). The amide functionality of almorexant forms a bidentate hydrogen bond with $N324^{6.55}$ and $H350^{7.39}$ (Fig. 3b), and mutation of the latter residue to Ala was shown experimentally to reduce binding affinity for hOX_2R[18]. In the predicted pose for EMPA, hydrogen bonds are established between the methoxy substituent on the 2-methoxypyridine and $T111^{2.61}$ and $Y354^{7.43}$ on the receptor (Fig. 3c), both of which are implicated in EMPA's interaction by mutational data[18,30]. In contrast to the other two molecules, no EMPA pose featured intramolecular π-stacking similar to suvorexant. For almorexant and EMPA, docking also yielded favourably scored second binding modes consistent with mutational data (Extended Data Fig. 6a, b). Finally, the predicted pose for SB-674042 closely resembles the binding mode of suvorexant, with its phenyl-oxadiazole overlapping almost perfectly with suvorexant's benzoxazole and its 2-methyl-thiazole overlapping with suvorexant's triazole (Fig. 3d). Overall, the prediction of intramolecular π-stacked conformations for multiple docked orexin receptor antagonists suggests that this property may be a general favourable design feature for synthetic molecules targeting the orthosteric site of hOX_2R.

We solved a high-resolution crystal structure of hOX_2R bound to the therapeutic compound suvorexant, providing a molecular framework for understanding DORA binding and stabilization of the inactive state by a salt-bridge network at the extracellular surface. Docking calculations predict putative stable binding modes for other orexin receptor antagonists, which are consistent with known mutational data. This knowledge will serve as a powerful tool in the design of improved agents that can activate or inactivate orexin signalling.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 16 September; accepted 4 November 2014.
Published online 22 December 2014.

1. Li, J., Hu, Z. & de Lecea, L. The hypocretins/orexins: integrators of multiple physiological functions. *Br. J. Pharmacol.* **171,** 332–350 (2014).

2. Michelson, D. *et al.* Safety and efficacy of suvorexant during 1-year treatment of insomnia with subsequent abrupt treatment discontinuation: a phase 3 randomised, double-blind, placebo-controlled trial. *Lancet Neurol.* **13,** 461–471 (2014).
3. Fredriksson, R., Lagerström, M. C., Lundin, L.-G. & Schiöth, H. B. The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. *Mol. Pharmacol.* **63,** 1256–1272 (2003).
4. Winrow, C. J. & Renger, J. J. Discovery and development of orexin receptor antagonists as therapeutics for insomnia. *Br. J. Pharmacol.* **171,** 283–293 (2014).
5. Zhu, Y. *et al.* Orexin receptor type-1 couples exclusively to pertussis toxin-insensitive G-proteins, while orexin receptor type-2 couples to both pertussis toxin-sensitive and -insensitive G-proteins. *J. Pharmacol. Sci.* **92,** 259–266 (2003).
6. Lin, L. *et al.* The sleep disorder canine narcolepsy is caused by a mutation in the hypocretin (orexin) receptor 2 gene. *Cell* **98,** 365–376 (1999).
7. Chemelli, R. M. *et al.* Narcolepsy in orexin knockout mice: molecular genetics of sleep regulation. *Cell* **98,** 437–451 (1999).
8. Nishino, S., Ripley, B., Overeem, S., Lammers, G. J. & Mignot, E. Hypocretin (orexin) deficiency in human narcolepsy. *Lancet* **355,** 39–40 (2000).
9. Cox, C. D. *et al.* Discovery of the dual orexin receptor antagonist [(7*R*)-4-(5-chloro-1,3-benzoxazol-2-yl)-7-methyl-1,4-diazepan-1-yl][5-methyl-2-(2*H*-1,2,3-triazol-2-yl)phenyl]methanone (MK-4305) for the treatment of insomnia. *J. Med. Chem.* **53,** 5320–5332 (2010).
10. Rosenbaum, D. M. *et al.* GPCR engineering yields high-resolution structural insights into 2-adrenergic receptor function. *Science* **318,** 1266–1273 (2007).
11. Warne, T. *et al.* Structure of a β1-adrenergic G-protein-coupled receptor. *Nature* **454,** 486–491 (2008).
12. Caffrey, M. Crystallizing membrane proteins for structure determination: use of lipidic mesophases. *Annu. Rev. Biophys.* **38,** 29–51 (2009).
13. White, J. F. *et al.* Structure of the agonist-bound neurotensin receptor. *Nature* **490,** 508–513 (2012).
14. Egloff, P. *et al.* Structure of signaling-competent neurotensin receptor 1 obtained by directed evolution in *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **111,** E655–E662 (2014).
15. Horcajada, C., Guinovart, J. J., Fita, I. & Ferrer, J. C. Crystal structure of an archaeal glycogen synthase: insights into oligomerization and substrate binding of eukaryotic glycogen synthases. *J. Biol. Chem.* **281,** 2923–2931 (2006).
16. Manglik, A. *et al.* Crystal structure of the μ-opioid receptor bound to a morphinan antagonist. *Nature* **485,** 321–326 (2012).
17. Wu, B. *et al.* Structures of the CXCR4 chemokine GPCR with small-molecule and cyclic peptide antagonists. *Science* **330,** 1066–1071 (2010).
18. Malherbe, P. *et al.* Mapping the binding pocket of dual antagonist almorexant to human orexin 1 and orexin 2 receptors: comparison with the selective OX_1 antagonist SB-674042 and the selective OX_2 antagonist *N*-ethyl-2-[(6-methoxy-pyridin-3-yl)-(toluene-2-sulfonyl)-amino]-*N*-pyridin-3-ylmethyl-acetamide (EMPA). *Mol. Pharmacol.* **78,** 81–93 (2010).
19. Kruse, A. C. *et al.* Structure and dynamics of the M3 muscarinic acetylcholine receptor. *Nature* **482,** 552–556 (2012).
20. Ballesteros, J. A. Activation of the β_2-adrenergic receptor involves disruption of an ionic lock between the cytoplasmic ends of transmembrane segments 3 and 6. *J. Biol. Chem.* **276,** 29171–29177 (2001).
21. Bokoch, M. P. *et al.* Ligand-specific regulation of the extracellular surface of a G-protein-coupled receptor. *Nature* **463,** 108–112 (2010).
22. Rasmussen, S. G. F. *et al.* Structure of a nanobody-stabilized active state of the β_2 adrenoceptor. *Nature* **469,** 175–180 (2011).
23. Kruse, A. C. *et al.* Activation and allosteric modulation of a muscarinic acetylcholine receptor. *Nature* **504,** 101–106 (2013).
24. Cox, C. D. *et al.* Conformational analysis of *N*,*N*-disubstituted-1,4-diazepane orexin receptor antagonists and implications for receptor binding. *Bioorg. Med. Chem. Lett.* **19,** 2997–3001 (2009).
25. Sakurai, T. *et al.* Orexins and orexin receptors: a family of hypothalamic neuropeptides and G protein-coupled receptors that regulate feeding behavior. *Cell* **92,** 573–585 (1998).
26. Kolb, P. *et al.* Structure-based discovery of β_2-adrenergic receptor ligands. *Proc. Natl Acad. Sci. USA* **106,** 6843–6848 (2009).
27. Brisbare-Roch, C. *et al.* Promotion of sleep by targeting the orexin system in rats, dogs and humans. *Nature Med.* **13,** 150–155 (2007).
28. Malherbe, P. *et al.* Biochemical and behavioural characterization of EMPA, a novel high-affinity, selective antagonist for the OX_2 receptor. *Br. J. Pharmacol.* **156,** 1326–1341 (2009).
29. Langmead, C. J. *et al.* Characterisation of the binding of [^{3}H]-SB-674042, a novel nonpeptide antagonist, to the human orexin-1 receptor. *Br. J. Pharmacol.* **141,** 340–346 (2004).
30. Tran, D.-T. *et al.* Chimeric, mutant orexin receptors show key interactions between orexin receptors, peptides and antagonists. *Eur. J. Pharmacol.* **667,** 120–128 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements We acknowledge support from the Welch Foundation (I-1770 to D.M.R.), the Searle Scholars Program (D.M.R.), a Packard Foundation Fellowship (D.M.R.), an Emmy Noether Fellowship of the German Research Foundation (KO-4095/1-1 to P.K.) and COST Action GLISTEN (CM1207 to P.K.). We thank D. Borek and Z. Otwinowski for assistance with diffraction data processing. The National Institute of General Medical Sciences and National Cancer Institute Structural Biology Facility at the Advanced Photon Source is funded in whole or in part with federal funds from the National Cancer Institute (ACB-12002) and the National Institute of General Medical Sciences (AGM-12006).

Author Contributions J.Y. expressed, purified and crystallized the hOX_2R–PGS fusion protein, collected diffraction data, and solved the structure. J.C.M. performed computational docking experiments on synthetic orexin receptor antagonists. P.K. supervised and performed computational docking experiments. D.M.R. supervised the overall project, assisted with collection of diffraction data and wrote the manuscript. All authors discussed the results and commented on the manuscript.

Author Information Atomic coordinates and structure factors for the reported crystal structure have been deposited in the PDB under accession 4RNB. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.M.R. (dan.rosenbaum@utsouthwestern.edu).

METHODS

Cloning, expression and purification. A DNA fragment corresponding to residues 1–386 of hOX_2R was cloned into a modified pFastBac (Invitrogen) baculovirus expression vector with the haemagglutinin (HA) signal sequence followed by the Flag tag at the N terminus[31]. The 58 C-terminal (intracellular) amino acids of hOX_2R were omitted owing to the prediction that they are unstructured and do not comprise part of the 7TM bundle. The hOX_2R–PGS fusion protein construct was generated by substituting a synthetic DNA fragment containing the 196-amino-acid coding sequence of *P. abysii* glycogen synthase (PDB accession 2BFW)[15] for residues 255–293 in the hOX_2R ICL3 using an adapted Multi-Site Quickchange protocol (Stratagene). For purification, a deca-histidine tag was added at the C terminus. The resulting construct was transfected into Sf9 cells to produce a recombinant baculovirus with the Bac-to-Bac system (Invitrogen). Sf9 cultures were infected with recombinant baculovirus at a cell density of 3×10^6 per ml and 1 μM suvorexant was added to the media. Infected cells were grown for 48 h at 27 °C, and cells were harvested and stored at −80 °C for future use.

Sf9 cell membranes were lysed in a hypotonic buffer containing 10 mM Tris pH 7.5, 1 mM EDTA, 160 μg ml^{-1} benzamidine, 100 μg ml^{-1} leupeptin, 2 mg ml^{-1} iodoacetamide and 1 μM suvorexant (Selleck Chemicals). Lysed membranes were re-suspended and homogenized by dounce in a buffer containing 50 mM Tris pH 7.5, 500 mM NaCl, 1% (w/v) *n*-dodecyl-β-D-maltopyranoside (DDM; Anatrace), 0.2% sodium cholate, 0.2% cholesteryl hemi-succinate (CHS), 10% glycerol, 2 mg ml^{-1} iodoacetamide and 5 μM suvorexant. Solubilization proceeded for 1 h at 4 °C, followed by ultracentrifugation for 30 min at 100,000*g*. After centrifugation, the solubilized supernatant supplemented with 20 mM imidazole was incubated with Ni-NTA agarose beads (GE Healthcare) in batch-binding mode for 3 h at 4 °C. After binding, beads were washed with 15 column volumes of Ni-NTA buffer: 50 mM Tris pH 7.5, 500 mM NaCl, 0.1% DDM, 0.02% sodium cholate, 0.02% CHS, 5% glycerol, 50 mM imidazole and 5 μM suvorexant. Protein was eluted with 5 column volumes of Ni-NTA wash buffer with 200 mM imidazole. The eluate from nickel-affinity chromatography was supplemented with 2 mM calcium and loaded onto M1 anti-Flag affinity beads (Sigma). Detergent was exchanged on the M1 resin from DDM to 0.05% lauryl maltose neopentyl glycol (LMNG; Anatrace). Receptor was eluted from the M1 beads with 200 μg ml^{-1} Flag peptide plus 5 mM EDTA. To remove N-linked glycan from the receptor, PNGaseF (NEB) was added and the reaction was incubated at 4 °C overnight. Finally, protein was concentrated in a 100 kDa cut-off Vivaspin column (Sartorius) and run on a Superdex 200 size exclusion column (GE Healthcare). The purified protein displayed a single monodisperse peak in the size exclusion profile (Extended Data Fig. 1a), and was >95% pure as judged by SDS–PAGE gel electrophoresis (Extended Data Fig. 1b).

Crystallization. Purified receptor was concentrated to >30 mg ml^{-1} using a Vivaspin concentrator with a 100 kDa molecular weight cut-off (Sartorius) and subjected to crystallization by the *in meso* method[32]. The concentrated protein was reconstituted into a lipid mixture containing monoolein plus 10% (w/w) cholesterol (Sigma), where the protein solution:lipid mass ratio was 2:3. Receptor and lipid components were mixed at room temperature using a syringe mixing apparatus. Crystallization experiments were carried out in 96-well glass sandwich plates (Molecular Dimensions) by a Gryphon LCP crystallization robot (Art Robbins Instruments) using a 40 nl protein cubic phase overlaid with 800 nl precipitant solution. Crystallization plates were incubated at 20 °C and initial crystals appeared after 24 h in a precipitant condition consisting of 100 mM MES pH 6.0, 30% PEG 400, 200 mM sodium formate. Crystals matured to full size in 3 days. Improved crystals were obtained in a condition consisting of 100 mM sodium citrate pH 5.9, 31% PEG 400, 200 mM sodium formate, 3% 2,5-hexanediol (Extended Data Fig. 2). Crystals were cryo-protected by harvesting directly from the LCP/precipitant set-ups with 100 μm MiTeGen loops and flash freezing in liquid nitrogen.

Data collection and processing. All diffraction data were collected at the 23ID-D beamline (GM/CA-CAT) at the Advanced Photon Source, Argonne National Laboratory, which is equipped with a Pilatus3 6M detector. Data sets were acquired using a 20 μm collimated minibeam with 1.033 Å wavelength X-rays. For a typical crystal, twenty-five 0.4° oscillation images were collected, with 1 s exposure and without attenuation of the beam, before radiation damage became excessive. Diffraction data from 52 crystals were merged into one complete data set. The resolution limit was set at 2.5 Å after anisotropy analysis with HKL3000 (ref. 33) (Extended Data Table 1).

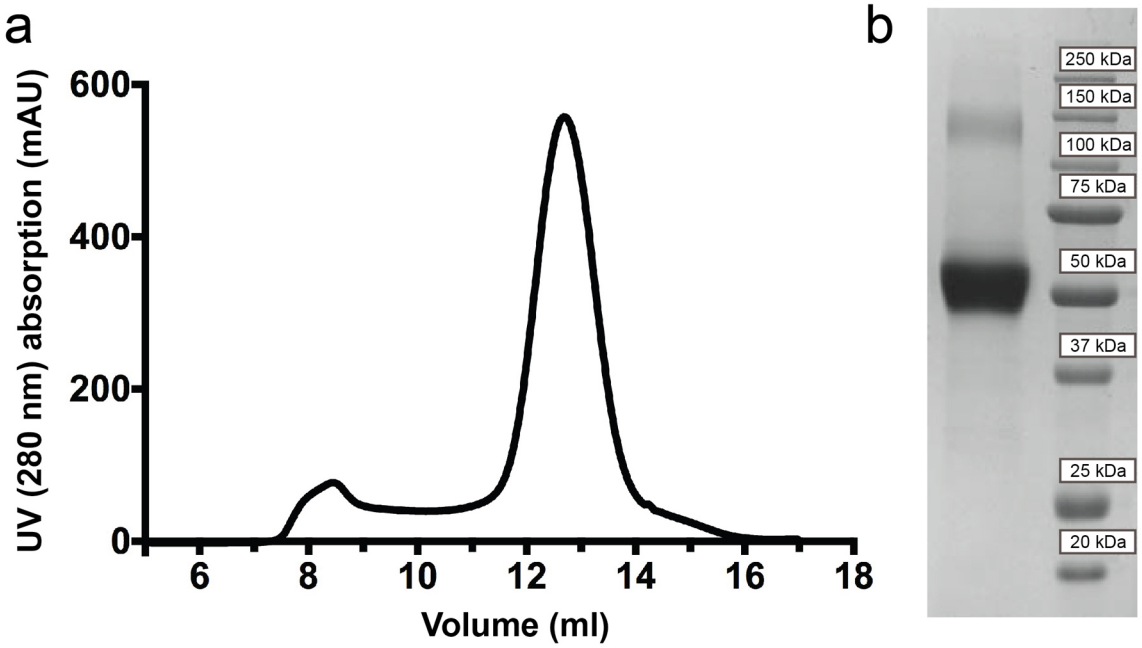
Structure determination and refinement. The structure of hOX_2R–PGS was solved by molecular replacement with Phaser[34] in Phenix[35]. The PGS domain (PDB accession 2BFW)[15] and μ-OR (PDB accession 4DKL)[16] were used as independent search models after analysis with Sculptor in Phenix[35]. The resulting solution was improved by auto-building in Buccaneer[36] and by manual iterative building in Coot[37] followed by refinement with Phenix. Translation–libration–screw (TLS) refinement was employed to model atomic displacement factors, with TLS groups generated by the TLSMD web server[38]. Initial coordinates and refinement parameters for the suvorexant ligand were prepared with the PRODRG[39] web server. An elongated feature in the electron density map, which was observed within the bilayer region, was modelled as oleic acid. MolProbity[40] was used to evaluate the final structure. In the Ramachandran plot, 98.1% of residues were in favoured regions and 1.9% of residues were in allowed regions. The statistics for data collection and refinement are included in Extended Data Table 1. Figures were prepared using PyMol (Schrodinger LLC). The electrostatic potential surface shown in Fig. 2a was calculated using DelPhi[41], and the ligand contact map shown in Fig. 2b was made using LIGPLOT[42].

Small-molecule docking. Docking calculations were done with DOCK 3.6 (refs 43, 44) and AutoDock[45] in order to obtain more diverse solutions. Dockings of the three orexin receptor antagonists to hOX_2R with AutoDock v.4.2 (ref. 45) used a static receptor and a flexible ligand. Receptor and ligand preparation was performed with Autodock Tools (ADT). The reference grid box ($60 \times 60 \times 60$ points and 0.375 Å of grid spacing) surrounded the suvorexant pose in the hOX_2R structure, allowing free ligand rotation and displacement. A genetic algorithm was used for exhaustive conformational sampling, and run 100 times with different random seeds.

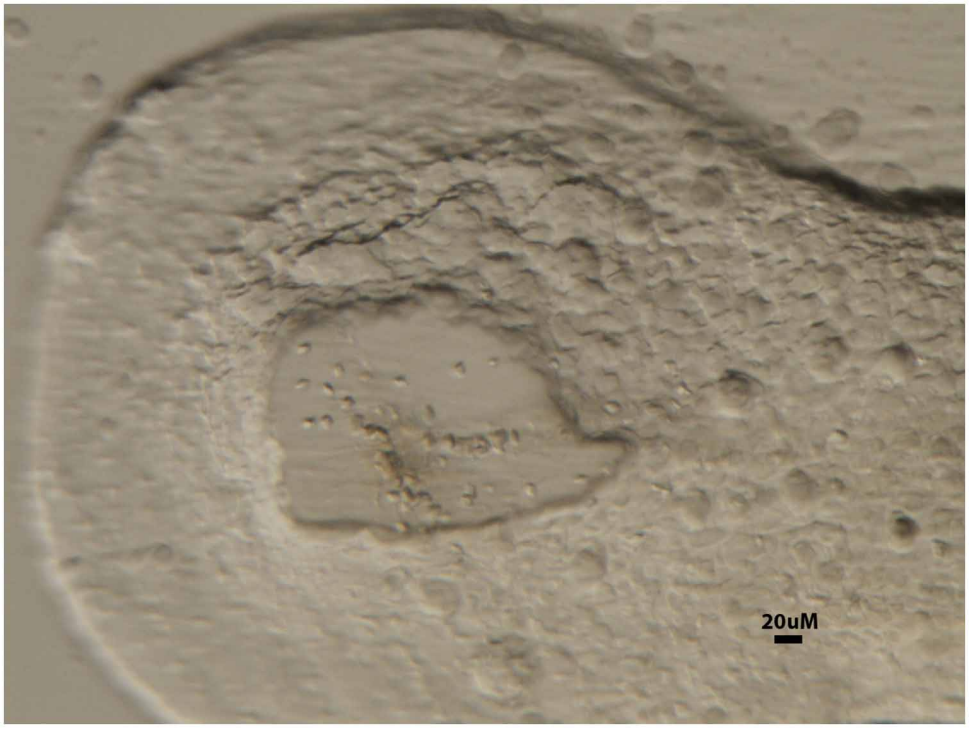
Docking of all compounds was also performed with DOCK 3.6 (refs 43, 44). Anchor spheres to guide the placement of the molecules were distributed based on the molecular surface of the receptor and the pose of suvorexant in the hOX_2R structure. The receptor was fixed during calculations and prepared for docking such that ionizable side chains were charged, except for histidines, for which protonation was modelled based on protein environment.

To further enrich conformational space, small-molecule conformations were generated with OMEGA[46] (OpenEye), using default settings except for the forcefield (mmff94s); an increased maximum number of conformations (300); an enlarged energy window (20); and a decreased r.m.s.d. cut-off (0.3). Representative conformations were then manually positioned in the binding pocket and minimized using the CHARMm22 forcefield (Accelrys), not constraining $N324^{6.55}$ and $H350^{7.39}$ to allow for side-chain flips. Expert criteria, namely satisfaction of hydrogen bonds, matching of polar and apolar groups, and consistency with mutational data, were used in the inspection and final selection of the poses. Finally, all poses, including the DOCK- and AutoDock-derived ones were evaluated with the DSX scoring function[47]. The poses shown were among the ones with the most favourable interaction scores. Two-dimensional chemical structures were drawn with Marvin 6.2.0 (Chemaxon).

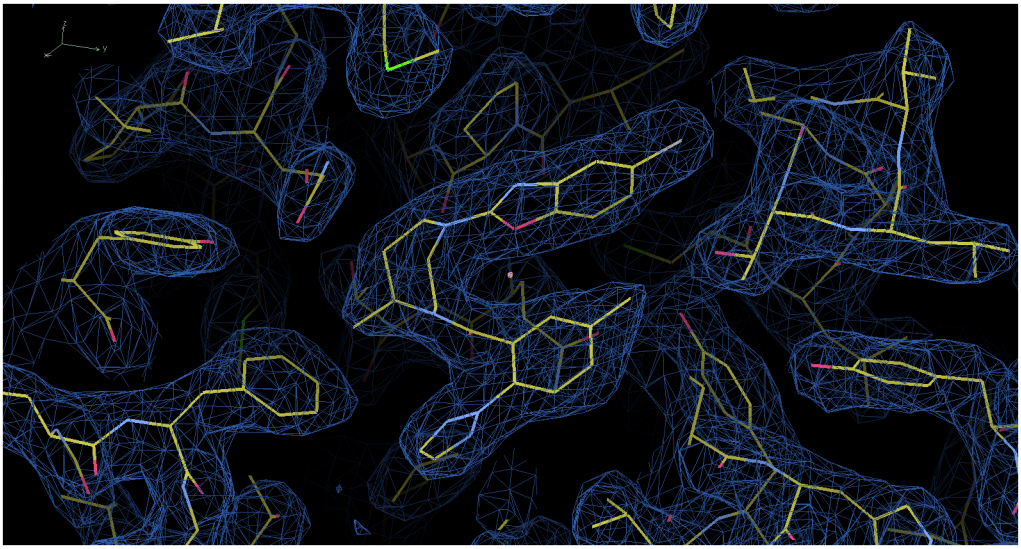
31. Kobilka, B. K. Amino and carboxyl terminal modifications to facilitate the production and purification of a G protein-coupled receptor. *Anal. Biochem.* **231,** 269–271 (1995).
32. Caffrey, M. & Cherezov, V. Crystallizing membrane proteins using lipidic mesophases. *Nature Protocols* **4,** 706–731 (2009).
33. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276,** 307–326 (1997).
34. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40,** 658–674 (2007).
35. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66,** 213–221 (2010).
36. Cowtan, K. Fitting molecular fragments into electron density. *Acta Crystallogr. D* **64,** 83–89 (2008).
37. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66,** 486–501 (2010).
38. Painter, J. & Merritt, E. A. TLSMD web server for the generation of multi-group TLS models. *J. Appl. Crystallogr.* **29,** 109–111 (2006).
39. Schüttelkopf, A. W. & van Aalten, D. M. F. PRODRG: a tool for high-throughput crystallography of protein-ligand complexes. *Acta Crystallogr. D* **60,** 1355–1363 (2004).
40. Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66,** 12–21 (2010).
41. Rocchia, W., Alexov, E. & Honig, B. Extending the applicability of the nonlinear Poisson–Boltzmann equation: multiple dielectric constants and multivalent ions. *J. Phys. Chem. B* **105,** 6507–6514 (2001).
42. Wallace, A. C., Laskowski, R. A. & Thornton, J. M. LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein Eng.* **8,** 127–134 (1995).
43. Irwin, J. J. *et al.* Automated docking screens: a feasibility study. *J. Med. Chem.* **52,** 5712–5720 (2009).
44. Mysinger, M. M. & Shoichet, B. K. Rapid context-dependent ligand desolvation in molecular docking. *J. Chem. Inf. Model.* **50,** 1561–1573 (2010).
45. Morris, G. M. *et al.* AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. *J. Comput. Chem.* **30,** 2785–2791 (2009).
46. Kirchmair, J., Wolber, G., Laggner, C. & Langer, T. Comparative performance assessment of the conformational model generators Omega and Catalyst: a large-scale survey on the retrieval of protein-bound ligand conformations. *J. Chem. Inf. Model.* **46,** 1848–1861 (2006).
47. Neudert, G. & Klebe, G. *DSX*: a knowledge-based scoring function for the assessment of protein-ligand complexes. *J. Chem. Inf. Model.* **51,** 2731–2745 (2011).



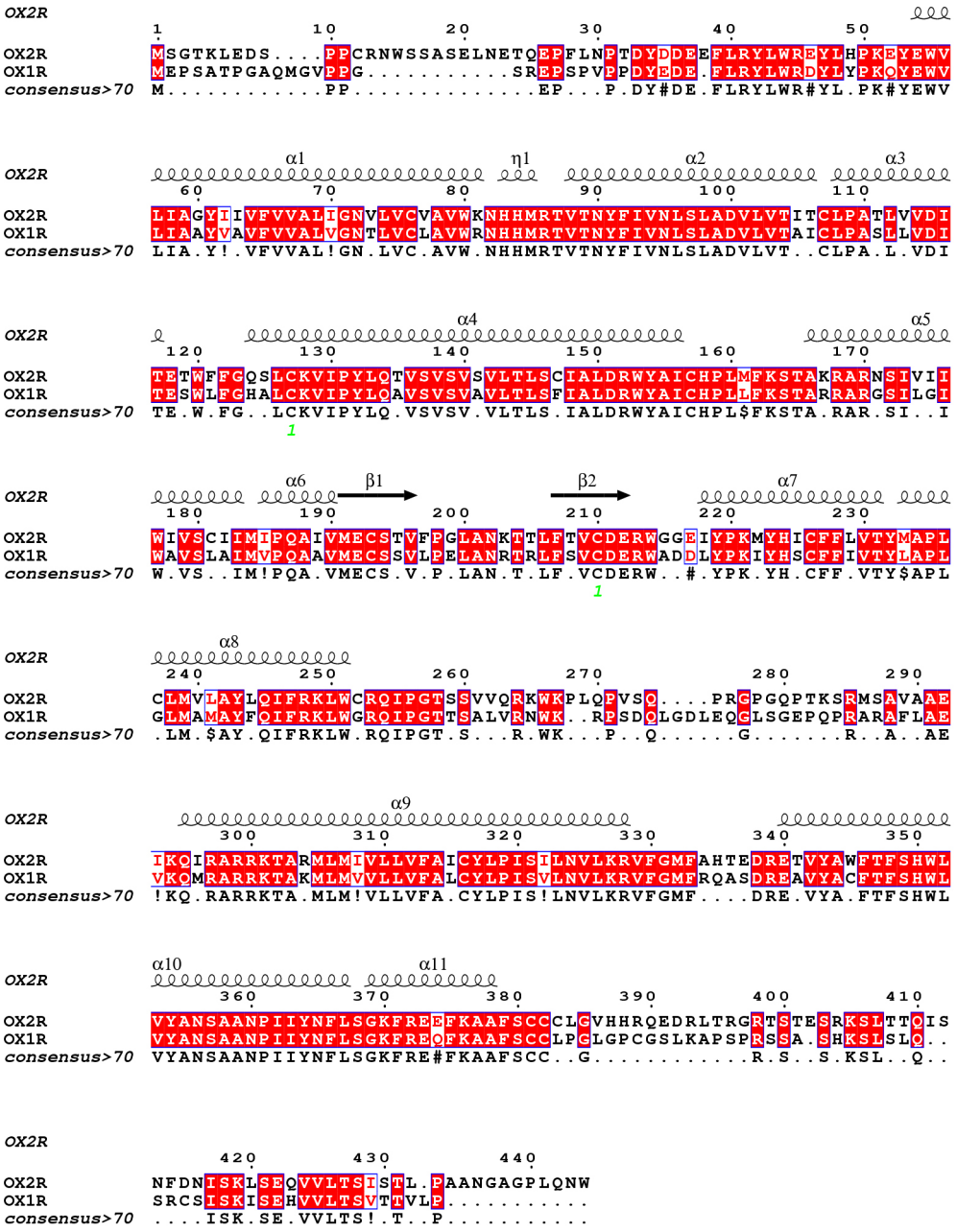
Extended Data Figure 1 | Purification of crystallization-grade hOX$_2$R–PGS. **a**, Superdex 200 gel filtration profile of hOX$_2$R–PGS purified by nickel immobilized-metal affinity chromatography (Ni-IMAC) and M1-Flag immunoaffinity chromatography. **b**, Coommassie-stained polyacrylamide gel electrophoresis (PAGE) of the isolated peak fraction from gel filtration.



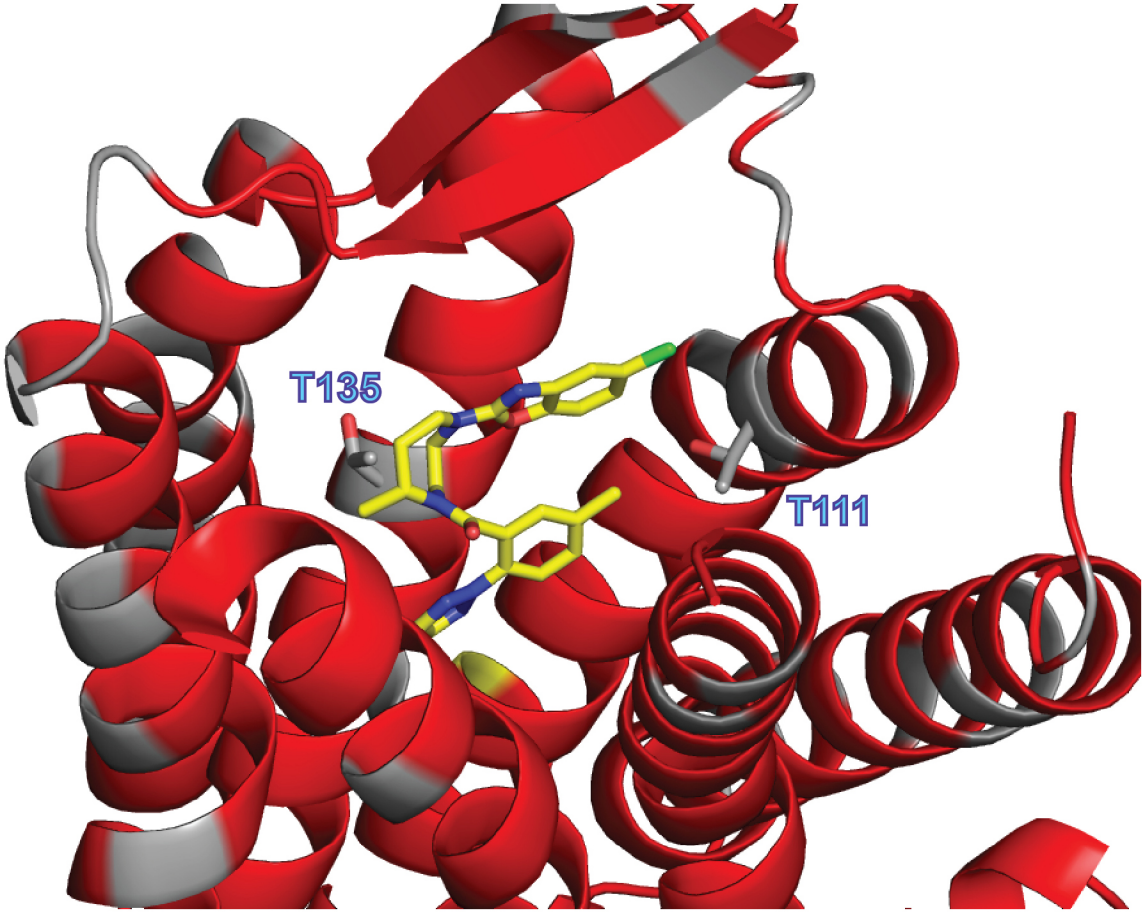
Extended Data Figure 2 | Lipidic cubic phase crystallization setup for hOX$_2$R–PGS. The image shows representative microcrystals of the hOX$_2$R–PGS protein that were harvested to produce high-resolution diffraction.



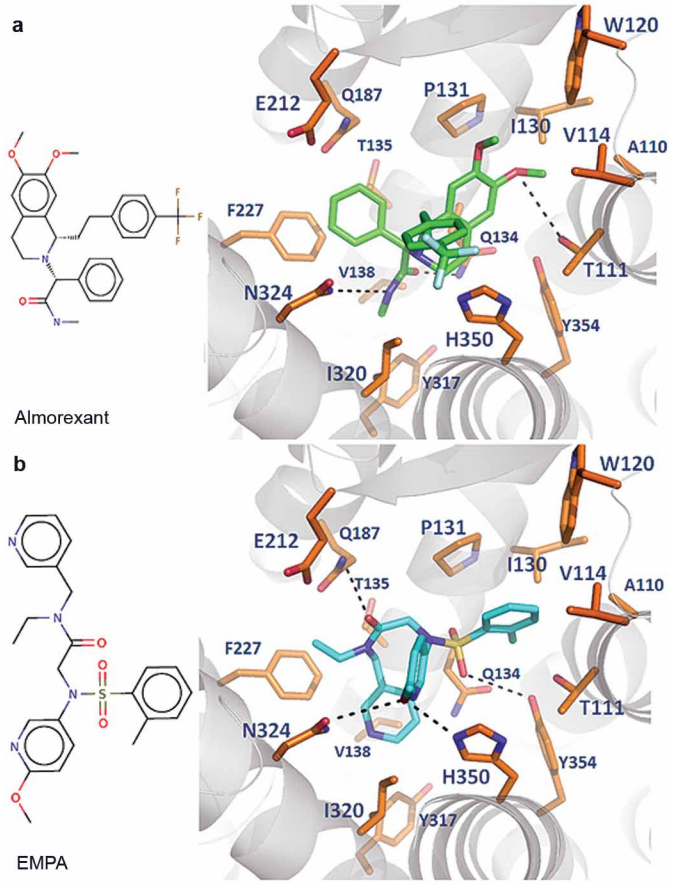
Extended Data Figure 3 | Electron density map for suvorexant and surrounding residues. The $2F_o - F_c$ electron density map is contoured at 1.2σ.



Extended Data Figure 4 | Sequence alignment between hOX$_2$R and hOX$_1$R. Positions that are identical between the two receptors are highlighted with a red background.



Extended Data Figure 5 | Conservation of the orthosteric binding pocket between hOX_2R and hOX_1R. Structure of the extracellular region of hOX_2R, with residues that are identical between hOX_2R and hOX_1R coloured red, and residues that are different coloured grey. $T111^{2.61}$ (to Ser) and $T135^{3.33}$ (to Ala) are the only residues within 6 Å of suvorexant that are different between the two GPCRs. ECL3 is removed for clarity.



Extended Data Figure 6 | Alternative docked poses for almorexant and EMPA. **a**, Left, chemical structure of almorexant. Right, second docked pose of almorexant (green carbons) that was favourably scored and in agreement with mutational data. **b**, Left, chemical structure of EMPA. Right, second docked pose of EMPA (cyan carbons) that was favourably scored and in agreement with mutational data.

Extended Data Table 1 | Data collection and refinement statistics

	hOX_2R-PGS*
Data collection	
Space group	C2
Cell dimensions	
a, *b*, *c* (Å)	94.36, 75.82, 96.30
α, β, γ (°)	90.00, 111.71, 90.00
Resolution (Å)	50.00(2.50) †
R_{sym} or R_{merge} ‡	0.21(N/A)
$I/\sigma I$ §	10.90/(0.86)
	*a**, (0.26)
	*b**, (2.00)
	*c**, (3.80)
Completeness (%)	99.90(99.00)
Redundancy	14.30(5.9)
Refinement	
Resolution (Å)	43.70-2.50 (2.6-2.50)
No. reflections	18,772
R_{work}/ R_{free}	0.19/0.24 (0.26/0.31)
No. atoms	
Protein	3,810
Ligand/ion	32
Water	36
B-factors	
Receptor	42.40
Fusion protein	48.90
Ligand/ion	26.90
Other (Lipid and water)	39.35
R.m.s deviations	
Bond lengths (Å)	0.004
Bond angles (°)	0.77

* Diffraction data from 52 crystals were merged into a complete data set.
† Highest-resolution shell is shown in parenthesis.
‡ R_{merge} higher than 1 is statistically meaningless, therefore Scalepack (HKL3000, ref. 33) does not report it.
§ Crystals diffracted anisotropically. The correction for anisotropy was applied during scaling with Scalepack (HKL3000). $I/\sigma I$ values (a*, b* and c*) for the highest-resolution shell (2.62–2.5 Å) were calculated by dividing mean intensity values in each direction with average error values.

WORKFORCE TRACKING Software tool automates data collection **p.253**

SATISFACTION Bad jobs do not get better, but good jobs continue to improve **p.253**

FUNDING How to smooth out spending bumps and busts **p.253**

RICHARD NOWITZ/GETTY

Young researchers may be especially vulnerable at field sites.

SOCIAL BEHAVIOUR

Indecent advances

Surveys of sexual harassment and assault during field research and on campus reveal a hitherto secret problem.

BY VIRGINIA GEWIN

Archaeologist Maureen Meyers never spoke up about the sexual harassment she endured from male colleagues and superiors at field sites and elsewhere during her 20-year career. She rebuffed a male colleague's propositions, leading to his retaliatory dismissal of her diabetes-related diet and medication requirements on a later field excursion. A male superior once forced her to walk ahead of him at a field site "to find the electric fences first" and made her listen to his lurid stories.

Meyers — now at the University of Mississippi in Oxford — considered abandoning her career several times. To help herself deal with what had happened, she recorded all her experiences in her diary. Last autumn, in response to the SAFE study documenting sexual harassment and assault in the field (K. B. H. Clancy *et al. PLoS ONE* **9,** e102172; 2014), Meyers organized a survey of archaeologists in the southeastern United States and learned the extent and severity of similar behaviour today. The responses to both surveys confirmed that she had been far from alone. And she counts herself lucky. "I was never physically assaulted," she says.

Many women who work in scientific disciplines involving remote fieldwork have experienced similar ordeals. But accounts of predatory behaviour have largely remained shrouded in secrecy, conveyed mostly as whispered warnings. Early-career researchers — mainly women, although men note harassment as well — are most vulnerable, yet are loath to speak up about sexual harassment, and even assault, lest their reputations be tainted and their careers damaged as a result of peer scepticism or retaliation by the offender.

According to the United Nations, harassment is defined as unwelcome sexual advances, requests for sexual favours and other verbal or physical conduct of an intimate nature. The definition comprises any such behaviour that creates a hostile or offensive work environment and can include hanging around the victim, making unwanted leading remarks or touching them, as well as attempted or actual sexual assault.

Studies also show that sexual harassment is usually more about power than about sex, making harassment by senior scientists of their subordinates the most difficult to deal with. Those who could be vulnerable to sexual harassment need to look to their personal safety — yet also have the problem of protecting their career.

Although still dismaying, the outlook may slowly be improving. Early-career researchers, both men and women, and academic organizations are beginning to develop individual and collective ways to protect potential and actual victims and to raise people's awareness that harassment and assault continues and how best to handle it, whether as victim or colleague.

SHINING A LIGHT

The incidence of sexual harassment and assault at scientific field sites was quantified last July when the SAFE study was published. This online survey of field scientists uncovered a range of negative experiences; nearly two-thirds of the 666 respondents, who were mostly women, reported being sexually harassed at a field site, and one-fifth said that they had been sexually assaulted. The findings stunned the scientific community and prompted dozens of news articles and thousands of social-media postings.

Those findings have sparked more surveys that will become the basis for clear guidelines on acceptable behaviour at field sites and reporting procedures. In Meyers' survey, for example, conducted at the behest of the Southeastern Archaeological Conference (SEAC), more than two-thirds of the almost 600 respondents said that they had experienced sexual harassment at a field site. ▶

Some 13% said that the harassment directly affected their careers, forcing them to change field sites, jobs or research interests, or to leave the discipline altogether.

And more than one-quarter said that the harassment had stymied their careers in other ways, such as causing them to question their abilities and their future in the discipline, fearing for their safety at field sites and being reluctant to conduct field research.

Pat Knezek, now a science administrator, worked as an astronomer for more than 20 years. When she was a junior researcher, magazine centrefolds were blatantly displayed at some US observatories, she says. That is not acceptable now, but more subtle predatory behaviour, such as invitations to junior researchers to discuss career prospects one-on-one after hours, continues and is harder to fight because it is less overt, she says.

Harassment at field sites is not the only problem. A slew of high-profile cases at US universities in the past few years has prompted federal directives that instruct universities to better respond to — and prevent — sexual assault on campus. As a result, there has been more attention to Title IX, the US federal law that prohibits sex discrimination (including sexual harassment or assault) on campus. More universities are forming offices that address the response to and prevention of sexual harassment and violence, says Joan Slavin, director of Northwestern University's Office of Sexual Harassment Prevention in Evanston, Illinois. And some professional scientific societies are creating guidelines and policies to deter predatory behaviour and to provide resources for female researchers who have been harassed or assaulted.

Most other countries have yet to catch up with the United States. Nicole Westmarland, co-director of the Centre for Research into Violence and Abuse at Durham University, UK, says that British efforts to stop harassment of women in academia are not at US levels. A letter that she co-authored in January in *The Telegraph* newspaper called for more clear-cut university policies on how to respond to sexual-assault complaints, and in an article in the newspaper a few days later she described UK universities' sexual-assault policies as "archaic". She says that university responses to sexual assault are most commonly described as inaction, either because sexual assault is a police matter beyond their remit or because they do not take disciplinary actions against the aggressor. Some Nordic universities are training employees to deal with sexual-harassment concerns, but many think that the issue is also under-studied there.

FIGHTING HARASSMENT

Some young researchers are getting creative. Upset by accounts of harassment at poster sessions or of fear of walking back to a hotel or campus after a conference party, two female postdocs have created a 'buddy system' called Astronomy Allies that they unveiled in January at an American Astronomical Society (AAS) meeting in Seattle, Washington. Participants volunteer to form a 'safe zone' — as a buffer, bystander or advocate — for AAS members who feel threatened or unsafe. In that case they can text or call an 'ally' for an escort.

Astronomy Allies has the support of the AAS Committee on the Status of Women in Astronomy. "We feel like we are breaking new ground by trying to make the community look at this issue — and find ways to protect the victims without putting ourselves in a position where we could get sued," says Joan Schmelz, committee chair and an astronomer at the University of Memphis in Tennessee, who herself was sexually harassed early in her career. For legal reasons, Schmelz can disclose no details, but she wrote in a 2011 blogpost that it involved both a sexual component and — as typifies such behaviour — an abuse of power.

Evidence of sexual harassment can help prevent the abuse, says anthropologist Kate Clancy.

L. BRIAN STAUFFER

"At the time, I was a young astronomer in a vulnerable position and the harasser was my supervisor," she wrote. She recalled that he told her that he wanted to put her in his pocket and take her out when it was convenient.

After that blogpost she became a go-to confidante for women grappling with similar experiences. Having heard many stories, she finds it difficult to offer general advice. "Rarely do I recommend filing an official report as a first action because it can affect your standing in your department and community — especially if you don't have a smoking-gun piece of evidence," she says. And publicly naming the harasser carries a risk of getting sued for defamation of character (see 'Anatomy of a sexual-harassment report'). Instead, she advises women to write down everything — the time, location, nature and details of an incident — and to save all evidence, including e-mails, texts and voice-mail messages. Then, she says, the victim should talk to someone they trust about the pros and cons of filing a report against the harasser (see 'What to do').

But she and others agree that predatory behaviour will stop only when the community decides that harassment and assault will not be tolerated and creates mechanisms that address them and make perpetrators accountable.

The issue has garnered less attention in scientific fields with greater gender parity, such as ecology. But Jacquelyn Gill, a palaeoecologist at the University of Maine in Orono, and Joshua Drew, a conservation ecologist at Columbia University in New York, will tackle the issue with a panel discussion at the August meeting of the Ecological Society of America in Baltimore, Maryland. "We want to start important conversations — for example, sharing university reporting procedures with students in their own labs, departments and institutions," says Gill. As a new principal investigator, she feels

BE PREPARED

What to do

Prevention tips

- Find out if there are rumours of sexual harassers in your field
- Familiarize yourself with the university's sexual-harassment policies and reporting protocols
- Discuss living arrangements and job expectations with your supervisor before going into the field
- Know whom to report sexual harassment concerns to while in the field
- Speak up if you see others in an uncomfortable, unsafe situation

How to respond to harrassment

- Save every correspondence (text, e-mail, voice mail, tweet) from the harasser
- Have witnesses to the harassment document what they saw
- Confide in a trusted colleague or friend and discuss the pros and cons of filing a report
- Contact your university's ombudsperson, Title IX representative, Human Resources Office, or Equal Employment Opportunity Office (any of these could trigger an investigation, however)
- Ask about university resources, including confidential counselling, no-contact orders issued by the university, workplace accommodations (schedule changes, office location changes, leave of absence), and referrals to advocates for legal, medical or housing assistance. V.G.

CASE STUDY

Anatomy of a sexual-harassment report

Sally Smith (not her real name) was a PhD student working at a remote marine field station in North America when a field-research supervisor propositioned her. When she turned down his advances, he threatened to bar her access to the gear and equipment that she needed to complete her fellowship research. Then came the domineering body language and verbal abuse.

She told the field-station manager, but he did nothing. Well-meaning senior women colleagues advised her not to draw attention to herself. Confused and vulnerable, she was unsure what to do, and ended up forgoing her fellowship, unwilling to put herself under his control for a second field season. But she received an alternative source of funding and continued her field work in the area — which led to more frightening encounters with him.

Smith wrote down every detail: dates, times and how the encounters made her feel. After her second field season, she took those records, along with every e-mail he had sent, to the ombudsman's office at her university. After she reported the harassment to the university's human-resources department, the perpetrator threatened to sue her for defamation. He ultimately lost his job, but later secured a post elsewhere, and Smith learned that he had continued to harass women.

"Unfortunately, speaking out is not always good for one's career, but it was worth the risk for me," she says. Now an assistant professor at a major university, Smith makes sure that her graduate students are prepared for safe, productive field experiences and know how to get help should they need it. That includes contacting her or the university ombudsman's office if they have intimidating encounters. **V.G.**

responsible for her graduate students. "We need to create a culture where incidents are rare and reporting is easy," she says.

CULTURAL SHIFT

The SAFE study is already starting to drive change. "While there have been anecdotes and whispers about harassment at field sites, scientists are trained to seek evidence in a methodical, quantitative way to confirm the presence of a problem," says Kate Clancy, a co-author of the SAFE paper and an anthropologist at the University of Illinois at Urbana-Champaign. "We gave them the data."

And SEAC past president Tristram Kidder, an anthropologist at Washington University in St Louis, Missouri, is helping to craft clear guidelines on professional field conduct and expectations as well as on detailed harassment-reporting procedures. They will be published this year. Other organizations in Europe and elsewhere are conducting discipline-based surveys in biology, astronomy, ecology and anthropology.

Some organizations, among them the American Geophysical Union, have already created a policy. The Association of American Geographers will draft guidelines for preventing and reporting harassment at its meeting in April, and the American Anthropological Association last year issued a 'zero tolerance' stance on sexual harassment and is launching an initiative to help members prevent it or deal with it when it happens.

Some groups are raising awareness through seminars. The online Earth Science Women's Network, an international peer-mentoring association, last autumn gave a presentation on field safety at the University of Wisconsin, Madison. "We talked about setting boundaries and expectations — about everything from living arrangements to working hours — before going into the field," says Erika Marin-Spiotta, a geographer at the university.

Others are working to change the culture of tacit acceptance nearer to home. Anthropologist Bob Muckle at Capilano University in Vancouver, Canada, says that he was stunned by the SAFE results. "I thought the stuff I had seen happen to female colleagues in the 1970s and 1980s had disappeared," he says. He has instituted a zero-tolerance policy on sexual harassment for the summer field school he directs, and gives students handouts that define harassment and provide contacts and phone numbers for reporting any such event.

Still, it will take more than lone actions or a few guidelines to effect a true cultural shift, say those who study the problem. Real change will come when the international scientific community decides, top-down and bottom-up, what constitutes acceptable behaviour. "Few things are simply a women's issue; this is a community issue," says SAFE co-author Julienne Rutherford, a biological anthropologist at the University of Illinois at Chicago. "Senior people in the hierarchy are more likely to be perpetrators. They are also the people who have the power to establish appropriate behaviour and what is acceptable in our work culture." ■

Virginia Gewin *is a freelance writer in Portland, Oregon.*

SOFTWARE

Career detective

Software that can track researchers' career progress is under development. It will automate the collection of data required to learn how and where young scientists get jobs. A team used data collected by the tool and by manual analysis to show that higher research output correlates with scientists' ability to move voluntarily between posts (A. Geuna *et al. Res. Policy* http://doi.org/2hz; 2015). Using researchers' names, the tool can mine web pages and CVs to identify affiliations and research productivity. The software could be used to reconstruct the career paths of researchers and to assess which factors are correlated with staying in academic positions or moving to another sector, says lead author Aldo Geuna, an economist at the University of Turin in Italy. The tool is openly available, he says, and developers and users are working to improve its algorithms.

EMPLOYMENT

Job dissatisfaction lasts

Women who dislike their job come to hate it more over time, even if they earn more, whereas men's job dissatisfaction stays much the same regardless of pay, according to a UK survey of 2,800 employees, which included scientists. Conversely, women and men who like their job enjoy it more as time passes. Kausik Chaudhuri at the University of Leeds, UK, a co-author of the study — called 'Job Satisfaction, Age and Tenure' — says the findings suggest that it does not become easier to adapt to a job that is not a good fit from the outset. Early-career researchers should therefore choose carefully in today's economic climate.

US RESEARCH FUNDING

Call to smooth bumps

Biomedical research advocates in the United States are calling for policy changes to ease boom-and-bust research-support cycles at the US National Institutes of Health (NIH). In a joint report, United for Medical Research, a research advocacy group, and the Information Technology & Innovation Foundation, a think tank in Washington DC, outline strategies to make the NIH budget more certain from year to year. These include apportioning federal funds for several years at a time and stipulating that any unspent funds can be rolled over to the next fiscal year.

THE EGG

All that remains.

BY S.B. DIVYA

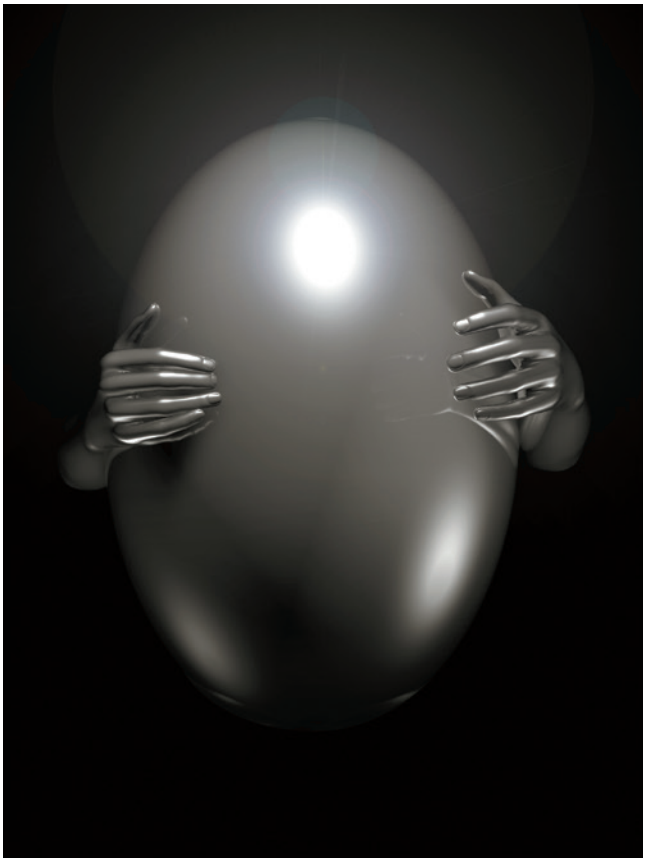


ILLUSTRATION BY JACEY

In the corner of the night-darkened room, tucked next to the sofa, the Egg rested on its pedestal like a modern sculpture. Its quiet hum was the only sound in the apartment; its green indicator the only light. The screen on the front of the ovoid was dark, not revealing the partially formed creature incubating within.

That wasn't right. The screen had never once been off, not while *she* had been here. She was gone now. She had slipped away quietly, without fuss, much as she'd lived.

"Promise," she had demanded, her voice raspy, as the smells of disinfectant and rot permeated his pores. "Promise that you'll keep it going."

"I promise," he'd lied. "Don't worry." He clutched the pills in his pocket with one hand.

In the end, she had been reduced to skin and bones. Her hand, clasping his, was a papery claw. She had always been scrawny. He'd called her chicken legs when they first met, and she'd retorted with "stupid head". Insults had never been her strong point. They were six years old. Love came years later, and the cancer not long after that.

She was cured the first time. A designer molecule flooded her system, keeping the traitorous cells at bay.

"Let's have a baby," she said when hope was allowed back into their house.

"Let's have two," he responded, and they both grinned like fools and got started.

They found out not long afterwards that the molecule that kept her alive was poison to any fetus. They spent the remainder of his inheritance on the Egg — and the hormones and extractions and fertilizations.

"It will be every bit your baby," promised the specialists.

She let them record her heartbeat and intestinal sounds for playback. The two of them used the microphone daily to stimulate budding ear drums. She sang her favourite songs in her off-key shower voice. He played his guitar and read cooking magazines aloud. They stared at the screen in fascination, watching it transform from a tadpole to an alien. The sofa seat nearest the Egg turned into a sinkhole.

The second cancer snuck in, quiet and efficient, while they were busy looking the other way. She needed another designer molecule, but she was too far down the queue. The money that would have bought her way higher was gone, so the doctors tried the old fashioned poisons. She lost her strength, the contents of her stomach and every hair on her body, but she didn't miss a day singing to the Egg.

Watching her reclining against the cylindrical pedestal, forehead resting on the warm ovoid above, he loved her even more.

"You're beautiful," he said.

She grinned, all teeth in a skeletal face. "You've never lied to me before."

"And I'm not lying now."

The second cancer took her swiftly. The apartment looked just as it had when they'd left for the hospital two days ago, but nothing was the same. The faint glow of city lights bled around the curtain edges, painting the room in a monochromatic palette. The Egg glinted, beckoning him. He shuffled towards it slowly like an old man and tripped on the edge of the rug — the rug that they'd chosen together to cushion tender baby feet and dimpled knees.

With a trembling hand he reached out and turned on the screen. It almost looked human now, although the head was too large and the body too skinny, sort of like she had looked in those last days of life. His hand moved of its own accord, navigating the menu screens, delving deep to find that buried option that came with every Egg. His fingers hovered over the number pad.

"I'm sorry, little one," he whispered. "This isn't how the road was supposed to go. I wish — if only —" He sighed. "I can't do this alone, and there's no one left for you but me, a poor excuse for a father." He drew his hand back. "Wait. Let's go together. I can do that much for you."

He stood and walked to the kitchen. His steps felt lighter now that the decision was made. He filled a glass with water, just enough to swallow a few pills. As he walked the scant distance back to the Egg, he reached into his pocket and retrieved the tablets. Their small white forms gleamed like pearls in his palm.

He reclined against the Egg, as she had, and closed his eyes. *You've never lied to me before.* Her words rattled like marbles in his skull. An involuntary tear traced its way down the contours of his face. It was the pinhole in the dam, and he felt all his grief push against it and then break through.

The sobs crashed over him in great waves, and he wrapped his arms around the warm Egg, clinging to it like a buoy in a storm. The glass and pills fell from his hands, forgotten in the tempest. An eternity passed before he went limp from exhaustion and fell asleep, his body curled around the Egg's pedestal. The menu system quietly and automatically exited to the start, and the screen went black. ■

S. B. Divya *holds degrees in computational neuroscience and signal processing. She is an engineer, the family calendar keeper and an author of short fiction. She also blogs at www.eff-words.com. This story is dedicated to the memory of her friend, Kevin Kanai Griffith (24 March–17 October 2014).*